ARMY RESEARCH LABORATORY

# Human Factors Experimental Design and Analysis Reference

**by Robert C. Willages**

**ARL-RP-0186** **July 2007**

A reprint from Virginia Polytechnic Institute and State University
Technical Report Number HFEEC-06-01 dated December 2006.

**Approved for public release; distribution is unlimited.**

## NOTICES

### Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturers' or trade names does not constitute an official endorsement or approval of the use thereof.

DESTRUCTION NOTICE—Destroy this report when it is no longer needed. Do not return it to the originator.

# Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5425

**ARL-RP-0186** **July 2007**


# Human Factors Experimental Design and Analysis Reference


**Robert C. Willages**
**Virginia Polytechnic Institute and State University**



A reprint from Virginia Polytechnic Institute and State University
Technical Report Number HFEEC-06-01 dated December 2006.



**Approved for public release; distribution is unlimited.**

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

**PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

| 1. REPORT DATE *(DD-MM-YYYY)* | 2. REPORT TYPE | 3. DATES COVERED (From - To) |
|---|---|---|
| July 2007 | Reprint | |

**4. TITLE AND SUBTITLE**

Human Factors Experimental Design and Analysis Reference

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**

Robert C. Willages (VPI)

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

U.S. Army Research Laboratory
Human Research & Engineering Directorate
Aberdeen Proving Ground, MD 21005-5425

**8. PERFORMING ORGANIZATION REPORT NUMBER**

ARL-RP-0186

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBERS**

**12. DISTRIBUTION/AVAILABILITY STATEMENT**

Approved for public release; distribution is unlimited.

**13. SUPPLEMENTARY NOTES**

The contracting officer's representative (COR) is Sam Middlebrooks, U.S. Army Research Laboratory, ATTN: AMSRD-ARL-HR-MV, Fort Hood, TX, telephone number 254-288-9379.

**14. ABSTRACT**

CADRE (computer-aided design reference for experiments) is a desktop computer tool for human factors and ergonomic researchers. The tool is in Acrobat[1] format for cross-platform computer use and has more than 500 bookmarks for information search. This tool provides more than 850 Power-Point[2] note pages of applied experimental design and analysis reference material. The reference material covers 25 topics and is divided into five major sections including introduction to experimental design, supplemental data collection design and analysis, basic analysis of variance (ANOVA) designs, advanced ANOVA designs, and empirical model building. References to the scientific literature are provided in each topic for supplemental reading. The CADRE tool also provides more than 200 pages explaining 40 examples of statistical analyses covered by the reference material. These examples are hyperlinked to Version 9.1.3 of the SAS[3] (2004) statistical analysis software, and the sample SAS programs can be easily modified for user-specific data.

[1]Acrobat is a registered trademark of Adobe Systems, Inc.
[2]PowerPoint is a trademark of Microsoft.
[3]SAS, which is not an acronym, is a registered trademark of the SAS Institute, Inc.

**15. SUBJECT TERMS**

experimental design

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | SAR | 1,130 | Robert C. Willages |
| UNCLASSIFIED | UNCLASSIFIED | UNCLASSIFIED | | | 19b. TELEPHONE NUMBER (*Include area code*) 540-951-7243 |

**Standard Form 298 (Rev. 8/98)**
**Prescribed by ANSI Std. Z39.18**

**Technical Report HFEEC-06-01**

# Human Factors Experimental Design and Analysis Reference

Robert C. Williges

**December 2006**

**Prepared For**

Collaborative Technology Alliance on Advanced Decision Architectures
Micro Analysis and Design
Colorado Springs, Colorado

and

U.S. Army Research Laboratory
Human Research and Engineering Directorate
Aberdeen Proving Ground, Maryland 21005-5425

**Human Factors Engineering and Ergonomics Center**
Grado Department of Industrial and Systems Engineering
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

## CADRE Description

CADRE is a desktop computer tool for human factors and ergonomic researchers. The tool is in PDF format for cross-platform computer use and has over 500 bookmarks for information search.

This tool provides over 850 PowerPoint note pages of applied experimental design and analysis reference material by Williges (2006). The reference material covers 25 topics and is divided into five major sections including introduction to experimental design, supplemental data collection design and analysis, basic analysis of variance (ANOVA) designs, advanced ANOVA designs, and empirical model building. References to the scientific literature are provided in each topic for supplemental reading,

The CADRE tool also provides over 200 pages in a companion appendix by Slater and Williges (2006) explaining 40 examples of statistical analyses covered by the reference material. These examples are hyperlinked to Version 9.1.3 of the SAS (2004) statistical analysis software, and the example SAS programs can be easily modified for user specific data.

## References

The CADRE project was supported by the Human Research and Engineering Directorate (HRED) of the Army Research Laboratory under the direction of Dr. Sam Middlebrooks and Dr. Michael Strub of HRED and through its prime contractor Micro Analysis and Design.

Slater, C.R. and Williges, R.C. (2006). Appendix: SAS examples for human factors experimental design and analysis reference. Blacksburg, VA: Virginia Polytechnic Institute and State University. Technical Report HFECC-06-02.

Williges, R.C. (2006). Human factors experimental design and analysis reference. Blacksburg, VA: Virginia Polytechnic Institute and State University. Technical Report HFECC-06-01.

Williges, R.C. (2007). CADRE: Computer-Aided Design Reference for Experiments (a tool for human factors and ergonomic research). Blacksburg, VA: Virginia Polytechnic Institute and State University. Electronic Book CD-ROM-07-01.

VirginiaTech
*Invent the Future*

# *CADRE*
# *Computer-Aided Design Reference for Experiments*

By
Robert C. Williges
R.H. Bogle Professor Emeritus
Grado Department of Industrial and Systems Engineering

2007

# CADRE: Computer-Aided Design Reference for Experiments (Williges, 2007)

## Topics in CADRE Material

Overview

Section 1
Introduction to Experimental Design

1. Research Design Process
2. Experimental Designs
3. Basic Statistical Concepts

Section 2
Supplemental Data Collection and Analysis

4. Supplemental Data Collection Methods
5. Analysis of Nominal Scale Data
6. Analysis of Ordinal Scale Data
7. Summary of Supplemental Data

Section 3
Basic Analysis of Variance (ANOVA) Designs

8. Introduction to ANOVA
9. ANOVA Summary Table
10. Between-Subjects ANOVA Designs
11. Analysis of Comparisons and Interactions
12. Within-Subjects ANOVA Designs
13. Mixed-Factors ANOVA Designs
14. Summary of Basic ANOVA

Section 4
Advanced ANOVA Designs

15. Introduction to Advanced ANOVA
16. Hierarchical ANOVA Designs
17. Blocking ANOVA Designs
18. Fractional-Factorial ANOVA Designs
19. Analysis of Covariance
20. Summary of Advanced ANOVA

Section 5
Empirical Model Building

21. Introduction to Empirical Models
22. Multiple Regression
23. Central-Composite Designs
24. Sequential Experimentation
25. Summary of Empirical Models

## What is CADRE?

- Applied Experimental Design and Analysis Reference
- A Computer Tool for Human Factors and Ergonomics Researchers

## Structure of CADRE

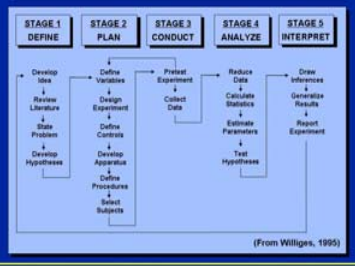
- Over 850 Pages of Reference Material
- Over 200 Pages of SAS Examples
- PDF Format for Cross Platform Compatibility

## Interactive Aspects of CADRE

- Over 500 PDF Bookmarks
- Links between Reference Material and Examples
- Hyperlinks to SAS Statistical Package

## Examples of CADRE Material

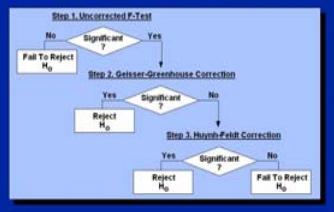
### 1.1. Stages of Research



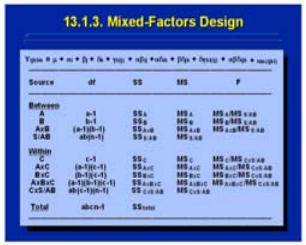
(From Williges, 1995)

### 12.2. Homogeneity of Covariance

- Strategy for Testing Main Effects and Interactions




### 13.1.3. Mixed-Factors Design

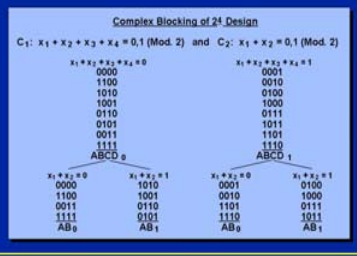


| Source | df | SS | MS | F |
|---|---|---|---|---|
| Between | | | | |
| A | a-1 | $SS_A$ | $MS_A$ | $MS_A/MS_{S:AB}$ |
| B | b-1 | $SS_B$ | $MS_B$ | $MS_B/MS_{S:AB}$ |
| AxB | (a-1)(b-1) | $SS_{AB}$ | $MS_{AB}$ | $MS_{AB}/MS_{S:AB}$ |
| S:AB | ab(n-1) | $SS_{S:AB}$ | $MS_{S:AB}$ | |
| Within | | | | |
| C | c-1 | $SS_C$ | $MS_C$ | $MS_C/MS_{CxS:AB}$ |
| AxC | (a-1)(c-1) | $SS_{AxC}$ | $MS_{AxC}$ | $MS_{AxC}/MS_{CxS:AB}$ |
| BxC | (b-1)(c-1) | $SS_{BxC}$ | $MS_{BxC}$ | $MS_{BxC}/MS_{CxS:AB}$ |
| AxBxC | (a-1)(b-1)(c-1) | $SS_{AxBxC}$ | $MS_{AxBxC}$ | $MS_{AxBxC}/MS_{CxS:AB}$ |
| CxS:AB | ab(c-1)(n-1) | $SS_{CxS:AB}$ | $MS_{CxS:AB}$ | |
| Total | abcn-1 | $SS_{total}$ | | |

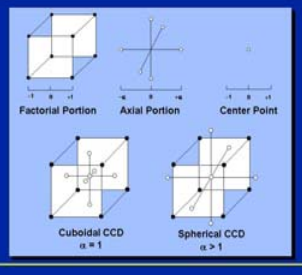
All the general procedures, rules, and algorithms apply to higher-order, mixed-factors designs. Always begin by stating the statistical model of the design. This slide shows the ANOVA Summary Table of a three-way, mixed-factors design that includes two between-subjects factors, A and B, and one within-subjects factor, C. The statistical model shows this designation in the nesting relationship for subjects, γ.

Notice that the between-subjects effects are grouped with their single error term, S/AB. Likewise, the within-subjects effects are grouped with their single error term, CxS/AB.

### 17.2.2. Complex Blocking of $2^k$ Design




### 23.2.3. Value of α in Central-Composite Design




## Content of CADRE Reference Material

- 1. Outline of Topic Contents
- 2. Applied Experimental Design and Analysis
  - Descriptions
  - Statistical Formulae and Algorithms
  - Experimental Design and Analysis Procedures
  - Guidelines
  - Calculation Procedures Linked to SAS Examples
  - References to Scientific Literature
- 3. Summary of Topic
- 4. Supplemental Readings

## CADRE Tool Extensions

- Additional Statistical Packages
- Family of Methodological Desktop Tools

## Content of SAS Examples in CADRE

- 1. Problem
  - Linked to CADRE Reference Material Calculations
  - Problem Description
  - Context/Purpose
  - Statistical Decision Criteria
- 2. SAS Input
  - SAS Program File Hyperlinked to SAS Statistical Package
- 3. SAS Output
  - Printout with Highlights and Notes
- 4. Output Explanation

## SAS Examples in CADRE

1. Interval Estimation
2. Single-Sample t-Test
3. Between-Subjects t-Test
4. Within-Subjects t-Test
5. Chi-Square Goodness of Fit Test
6. Chi-Square Test of Independence (2x2 Contingency Table)
7. Chi-Square Test of Independence (RxC Contingency Table)
8. Chi-Square Test of Independence (Two Additive 2x2 Partitions)
9. McNemar Change Test
10. Cochran Q Test
11. Kolmogorov-Smirnov Tests
12. Kruskal-Wallis One-Way ANOVA
13. Wilcoxon Signed Ranks Test
14. Friedman Two-Way ANOVA
15. One-Factor, Between-Subjects ANOVA
16. Two-Factor, Between-Subjects ANOVA
17. Planned Comparisons
18. Unplanned Comparisons
19. Analysis of Interactions
20. One-Factor, Within-Subjects ANOVA
21. Two-Factor, Within-Subjects ANOVA
22. Geisser-Greenhouse and Huynh-Feldt Corrections
23. Testing Order Effects in Balanced Latin Squares
24. ANOVA Design Comparison
25. Two-Way, Mixed-Factors ANOVA
26. Complete Hierarchical Between-Subjects Design
27. Partial Hierarchical Between-Subjects Design
28. Simple Blocking of $2^k$ Within-Subjects Design
29. Complex Blocking of $2^k$ Within-Subjects Design
30. One-Half Replicate of $2^4$ Between-Subjects Design
31. 4x4 Latin Square Designs
32. Linear Correlation Coefficient
33. Alternative Linear Correlations
34. Simple Linear Regression
35. One-Way Analysis of Covariance (ANCOVA)
36. Multiple Linear Regression
37. Best Regression Equation
38. Polynomial Regression
39. Orthogonal, Between-Subjects, Central-Composite Design
40. Blocked, Within-Subjects Central-Composite Design

## Table of Contents

**Topics** **Page**

**Overview**
0.1. Purpose of Reference Material ……3
0.1.1. Applied Experimental Design
0.1.2. Human Factors Engineering Methods
0.1.2.1. Research Methods
0.1.2.2. User Interface Design Methods
0.1.2.3. Training System Design Methods
0.1.2.4. Usability Evaluation Methods
0.2. Presentation Approach ……10
0.2.1. Format of Reference Material
0.2.2. Experimental Design References
0.3. Organization of Reference Topics ……13
0.3.1. Introduction to Experimental Design
0.3.2. Supplemental Data Collection and Analysis
0.3.3. Basic Analysis of Variance Designs
0.3.4. Advanced Experimental Designs

**Section 1. Introduction to Experimental Design**

Topic 1. Research Design Process
1.1. Stages of Research ……21
1.2. Research Problem ……22
1.3. Research Approach ……23
1.4. Critical Research Methods ……24
1.4.1. Variables
1.4.2. Procedures
1.4.3. Protection of Human Subjects
1.4.4. Equipment
1.4.5. Pretesting
1.5. Research Design Alternatives ……30
1.6. Analyzing Results ……31
1.7. Research Reports ……32
1.7.1. Scientific Reports
1.7.2. Major Components of Reports
1.7.2.1. Introduction
1.7.2.2. Method
1.7.2.3. Results
1.7.2.4. Discussion
1.7.3. Additional Considerations
1.8. Summary ……41
1.9. Supplemental Readings ……44

Topic 2. Experimental Designs
2.1. Introduction ……46
2.1.1. Threats to Validity
2.1.2. Quantitative Research Approach
2.2. Experimental Design Alternatives ……51
2.2.1. Experimental Design Notation
2.2.2. Quasi-Experimental Designs
2.2.2.1. Single Group Designs
2.2.2.2. Nonequivalent Control Group Designs
2.2.2.3. Regression-Discontinuity Designs
2.2.2.4. Interrupted Time Series Designs
2.2.3. Randomized Experimental Designs
2.3. Summary ……69
2.4. Supplemental Readings ……70

Topic 3. Basic Statistical Concepts
3.1. Probability ……72
3.1.1. Compositional Techniques
3.1.2. Counting Techniques
3.2. Random Sampling ……81
3.3. Sampling Distributions ……82
3.3.1. Binomial Distribution
3.3.2. Normal Distribution
3.3.3. Student's t Distribution
3.3.4. Chi-Squared Distribution
3.3.5. F-Distribution
3.4. Statistical Estimation ……99
3.4.1. Estimators
3.4.2. Point Estimation
3.4.3. Interval Estimation
3.4.4. Summary of Statistical Estimation
3.5. Statistical Hypothesis Testing ……109
3.5.1. Components of Hypothesis Testing
3.5.1.1. Null and Alternative Hypotheses
3.5.1.2. Format for Hypothesis Test
3.5.1.3. Types of Errors
3.5.1.4. Statistical Power
3.5.2. Single-Sample t-Test
3.5.3. Relationship to Statistical Estimation
3.6. Two-Sample t-Tests ……122
3.6.1. Sampling Distribution
3.6.2. Assumptions
3.6.3. Standard Format
3.6.4. Between-Subjects t-Test
3.6.5. Within-Subjects t-Test
3.6.6. Conclusion

3.7. Summary ……..139
3.8. Supplemental Readings ……..140

**Section 2. Supplemental Data Collection and Analysis**

Topic 4. Supplemental Data Collection Methods
1.1. Background ……..143
1.1.1. Types of Dependent Variables
1.1.2. Analysis Procedures
1.2. Nonparametric Procedures ……..147
1.2.1. Scales of Measurement
1.2.2. Classification Scheme
1.3. Subjective Measures ……..154
1.3.1. Self Reports
1.3.1.1. Verbal Protocols
1.3.1.2. Critical Incidents
1.3.2. Questionnaires
1.3.3. Psychometric Scaling
1.4. Graphic Rating Scales ……..164
1.4.1. Likert Rating Scales
1.4.2. Bipolar Adjective Scales
1.4.3. Rating Scale Reliability and Validity
1.4.4. Examples of Rating Scales
1.5. Summary ……..177
1.6. Supplemental Readings ……..178

Topic 5. Analysis of Nominal Scale Data
5.1. Background ……..180
5.2. Between-Subjects Tests ……181
5.2.1. Chi-Square Goodness of Fit Test
5.2.2. Chi-Square Test of Independence
5.3. Within-Subjects Tests……..195
5.3.1. McNemar Change Test
5.3.2. Cochran Q Test
5.4. Summary ……..202
5.5. Supplemental Readings ……203

Topic 6. Analysis of Ordinal Scale Data
6.1. Background ……..205
6.2. Between-Subjects Tests ……206
6.2.1. Kolmogorov-Smirnov Tests
6.2.2. Kruskal-Wallis One-Way ANOVA
6.3. Within-Subjects Tests……..217
6.3.1. Wilcoxon Signed Ranks Test
6.3.2. Friedman Two-Way ANOVA
6.4. Summary ……..226

6.5. Supplemental Readings……..227

Topic 7. Summary of Supplemental Data
7.1. Supplemental Data Collection……229
7.1.1. Self Reports and Questionnaires
7.1.2. Rankings and Rating Scales
7.2. Supplemental Data Analysis……..232
7.2.1. Nominal Scale Data Analysis
7.2.2. Ordinal Scale Data Analysis
7.3. Supplemental Data Process……235
7.4. Summary……..238
7.5. Supplemental Readings……..239

**Section 3. Basic Analysis of Variance (ANOVA) Designs**

Topic 8. Introduction to ANOVA
8.1. Advantages of ANOVA Designs……...242
8.2. Basic Terms ……243
8.3. ANOVA Design Alternatives……...246
8.4. ANOVA Statistical Models ……....250
8.4.1. Specification Procedures
8.4.2. Examples
8.5. ANOVA Hypothesis Testing ..……258
8.5.1. Format of F-Test
8.5.1.1. Theoretical F
8.5.1.2. Hypotheses
8.5.1.3. Complete Format
8.5.2. Assumptions of the F-Test
8.5.3. Two-Level Design
8.5.3.1. Components of Deviation Score
8.5.3.2. Estimation of Population Variance
8.5.3.3. Hypothesis Test
8.6. Summary……272
8.7. Supplemental Readings……273

Topic 9. ANOVA Summary Table Components
9.1. Introduction ……275
9.2. Sources of Variation ……276
9.3. Degrees of Freedom (df)……280
9.3.1. Rules for Determining df
9.3.2. df Examples
9.4. Sum of Squares (SS) ……285
9.5. Mean Squares (MS)……286
9.5.1. Expected Mean Squares, E(MS)
9.5.2. Algorithm for Stating E(MS)
9.5.3. E(MS) Examples

9.6. F-Ratios ……297
9.6.1. Rules for Determining F-Ratios
9.6.2. F-Ratio Examples
9.7. Complete ANOVA Summary Table……...303
9.7.1. Summary Table Components
9.7.2. Summary Table Examples
9.8. ANOVA Design Construction ……308
9.9. Summary ……309
9.10. Supplemental Readings ……310

Topic 10. Between-Subjects ANOVA Designs
10.1. One-Factor, Between-Subjects Design……....312
10.1.1. One-Factor Design Example
10.1.2. Sum of Squares Calculations
10.1.2.1. Simplified Design Notation
10.1.2.2. SS Computational Formulae Algorithm
10.1.2.3. SS Numerical Computations
10.1.3. Summary Table and Test Format
10.2. Two-Factor, Between-Subjects Design ……...324
10.2.1. Design Configuration
10.2.2. AxB Interaction
10.2.3. Calculations
10.2.4. Two-Factor Design Example
10.3. n-Factor, Between-Subjects Design……..336
10.3.1. Three-Factor Design
10.3.2. Generalizations
10.4. Summary……340
10.5. Supplemental Readings ……341

Topic 11. Analysis of Comparisons and Interactions
11.1. Multiple Comparisons……...343
11.1.1. Linear Comparisons
11.1.2. Inflated Type I Error
11.1.3. Planned Comparisons
11.1.3.1. Planned F-Test
11.1.3.2. Critical Difference
11.1.3.3. Planned Bonferroni t Test (Dunn Test)
11.1.4. Unplanned Comparisons
11.1.4.1. Least Significant Difference Test (LSD)
11.1.4.2. Bonferroni t Test (Dunn Test)
11.1.4.3. Scheffé Multiple Contrast Procedure
11.1.4.4. Tukey's Honestly Significant Difference Test (HSD)
11.1.4.5. Dunnett Test
11.1.4.6. Newman-Keuls Sequential Range Test
11.1.4.7. Choice of Procedure
11.2. Evaluating Interactions……376

11.2.1. Example Problem
11.2.2. Graphing Procedures
11.2.3. Simple Effects Test
11.2.4. Trend Analysis
11.2.5. Paired Comparisons
11.2.5.1. Sequential Range Test
11.2.5.2. Unconfounded Comparisons
11.2.6. Interaction Evaluation Process
11.3. Summary……399
11.4. Supplemental Readings ……400

Topic 12. Within-Subjects ANOVA Designs
12.1. Within-Subjects Design Configurations……402
12.1.1. Single-Factor Design
12.1.2. Two-Factor Design
12.1.3. n-Factor Design
12.2. Homogeneity of Covariance ……417
12.3. Balancing Order of Treatments……423
12.3.1. Balancing Alternatives
12.3.2. Balanced Latin Square
12.3.3. Testing Order Effects
12.4. Differential Transfer……437
12.5. Within-Subjects Design Advantages ……438
12.6. Summary……440
12.7. Supplemental Readings ……441

Topic 13. Mixed-Factors ANOVA Designs
13.1. Mixed-Factors Design Configurations……443
13.1.1. Two-Factor Design
13.1.2. Two-Factor Design Example
13.1.3. Three-Factor Design
13.1.4. n-Factor Design
13.2. Mixed-Factors Design Considerations ……456
13.3. Summary……457
13.4. Supplemental Readings ……458

Topic 14. Summary of Basic ANOVA
14.1. Basic Considerations……460
14.2. ANOVA Rules and Algorithms ……461
14.2.1. Specification of Statistical Models
14.2.2. Rules for Degrees of Freedom
14.2.3. SS Computational Formulae Algorithm
14.2.4. Algorithm for Stating E(MS)
14.2.5. Steps for Determining F-Ratios
14.3. Design Classification…... ……467
14.4. n-Factor Design Generalizations……471

14.5. ANOVA Design Process……..472
14.6. Summary……...475
14.7. Supplemental Readings ……..476

**Section 4. Advanced ANOVA Designs**

Topic 15. Introduction to Advanced ANOVA
15.1. Basic ANOVA Extensions……...479
15.1.1. Quasi-F Ratios
15.1.2. Randomized Blocks Design
15.1.2.1. Constructing Randomized Blocks
15.1.2.2. Design Comparison
15.1.2.3. Extensions of Randomized Blocks
15.2. Advanced ANOVA Design and Analysis……...489
15.3. Summary……490
15.4. Supplemental Readings ……..491

Topic 16. Hierarchical ANOVA Designs
16.1. Basic Hierarchical Designs ……493
16.1.1. Between-Subjects Design
16.1.2. Within-Subjects Design
16.1.3. Mixed-Factors Design
16.2. Hierarchical Design Examples ……...503
16.2.1. Complete Hierarchical Design
16.2.2. Partial Hierarchical Design
16.3. Summary……516
16.4. Supplemental Readings ……..517

Topic 17. Blocking ANOVA Designs
17.1. Modular Representation……..519
17.1.1. Modular Arithmetic
17.1.2. Balanced Sets of Treatments
17.1.3. Component SS Formulae
17.1.4. Generalizations
17.2. Blocking $2^k$ Designs ……533
17.2.1. Simple Blocking of $2^k$ Design
17.2.2. Complex Blocking of $2^k$ Design
17.2.3. Computational Considerations
17.2.3.1. Simple Blocking Example
17.2.3.2. Complex Blocking Example
17.3. Pseudo-Factor Blocking ……..554
17.4. Summary……558
17.5. Supplemental Readings ……..559

Topic 18. Fractional-Factorial ANOVA Designs
18.1. 2k-p Fractional Replicates ……..561

18.1.1. Design Construction
18.1.1.1. One-Half Replicate
18.1.1.2. One-Fourth Replicate
18.1.2. Computational Considerations
18.1.3. Design Resolution
18.1.3.1. Design Effects
18.1.3.2. Definitions of Resolution
18.1.3.3. Identity Relationship
18.1.3.4. Resolution III Design
18.1.3.5. Resolution IV Design
18.1.3.6. Resolution V Design
18.1.3.7. Uses of Design Resolution
18.2. Latin Square ANOVA Designs ……..592
18.2.1. Design Construction
18.2.1.1. Standard Latin Square
18.2.1.2. Balanced Latin Square
18.2.1.3. Relationship to Fractional Replicates
18.2.2. Computational Considerations
18.2.2.1. Additivity Assumption
18.2.2.2. Between-Subjects Design
18.2.2.3. Within-Subjects Design
18.2.2.4. Latin Square Examples
18.2.3. Design Constraints
18.3. Summary……611
18.4. Supplemental Readings ……..612

Topic 19. Analysis of Covariance (ANCOVA)
19.1. Introduction to ANCOVA……..614
19.2. Linear Correlation, $r_{12}$ ……..616
19.2.1. Correlation Coefficient
19.2.1.1. Computational Formulae
19.2.1.2. Tests of Significance
19.2.2. Alternative Correlations
19.2.2.1. Point Biserial Correlation, $r_{pbi}$
19.2.2.2. Phi Correlation, $r_{\phi}$
19.2.2.3. Spearman Correlation, $r_{\rho}$
19.2.2.4. Partial Correlation, $r_{(1.3)(2.3)}$
19.2.2.5. Semipartial Correlation, $r_{1(2.3)}$
19.3. Simple Linear Regression……..654
19.3.1. Line of Best Fit
19.3.1.1. Method of Least Squares
19.3.1.2. Calculation Example
19.3.1.3. Standardized Regression
19.3.2. Goodness of Fit
19.3.2.1. Partitioning Variance
19.3.2.2. Tests of Significance

19.3.2.3. Coefficient of Determination
19.4. ANCOVA Computations……..670
19.4.1. Basic ANCOVA Design
19.4.2. Advanced ANCOVA
19.4.3. Interpreting ANCOVA
19.5. Summary……685
19.6. Supplemental Readings ……..686

Topic 20. Summary of Advanced ANOVA
20.1. ANOVA Design Constraints……688
20.1.1. Random-Effects Factors
20.1.2. Nested Factors
20.1.3. Control of Nuisance Factors
20.1.4. Data Collection Limitations
20.1.5. Control of Subject Variability
20.2. Advanced ANOVA Design Process……...696
20.3. ANOVA of Regression Analysis……697
20.4. Summary……698
20.5. Supplemental Readings ……..699

**Section 5. Empirical Model Building**

Topic 21. Introduction to Empirical Models
21.1. Quantitative Models……..702
21.1.1. Mechanistic Models
21.1.2. Empirical Models
21.2. Model Building Experiments……...708
21.3. Models in Human Factors……...710
21.4. Summary……713
21.5. Supplemental Readings ……..714

Topic 22. Multiple Regression
22.1. Multiple Regression Procedures ……716
22.2. Multiple Linear Regression……..721
22.2.1. Line of Best Fit
22.2.2. Goodness of Fit
22.2.3. Multiple Regression Example
22.2.4. Best Regression Equation
22.2.4.1. Backward Selection
22.2.4.2. Forward Selection
22.2.4.3. Stepwise Selection
22.2.4.4. All Possible Regressions
22.2.5. Best Equation Example
22.3. Second-Order Polynomial Regression……..753
22.3.1. Polynomial Regression Calculations
22.3.2. Polynomial Regression Example

22.4. Summary……764
22.5. Supplemental Readings……765

Topic 23. Central-Composite Designs (CCD)
23.1. CCD Introduction……767
23.2. CCD Specification ……771
23.2.1. CCD Configuration
23.2.2. Replication
23.2.3. Value of $\alpha$
23.3. CCD Analysis……785
23.3.1. Between-Subjects CCD
23.3.2. Within-Subjects CCD
23.3.3. Mixed-Factors CCD
23.4. CCD Examples……794
23.4.1. Between-Subjects Example
23.4.2. Within-Subjects Example
23.5. Alternative Second-Order Designs ……806
23.6. Summary……809
23.7. Supplemental Readings ……810

Topic 24. Sequential Experimentation
24.1. Strategies for Experimentation……812
24.2. Response Surface Methodology (RSM)……816
24.2.1. Steps in RSM
24.2.2. Method of Steepest Ascent
24.3. Sequential Research……822
24.3.1. Sequential Research Paradigm
24.3.2. Sequential Research Example
24.3.3. Guidelines for Sequential Research
24.4. Integrated Research Database……849
24.5. Summary……852
24.6. Supplemental Readings ……853

Topic 25. Summary of Empirical Models
25.1. Model Building Experiments……855
25.2. Components of Empirical Models……856
25.3. Sequential Experimentation Process ……858
25.4. Overall Conclusions ……863
25.5. Summary……864
25.6. Supplemental Readings ……865

**Acknowledgements**……866

**References**……867

**Subject index**……874

## Notes on Computational Examples

**Examples** **Page**

Example 1. Interval Estimation ……..107
Example 2. Single-Sample t-Test ……..117
Example 3. Between-Subjects t-Test ……..130
Example 4. Within-Subject t-Test ……..136
Example 5. Chi-Square Goodness of Fit Test ……..184
Example 6. Chi-Square Test of Independence (2x2 Contingency Table) ……189
Example 7. Chi-Square Test of Independence (RxC Contingency Table)……192
Example 8. Chi-Square Test of Independence (Two Additive 2x2 Partitions) ……..193
Example 9. McNemar Change Test ……..198
Example 10. Cochran Q Test ……..200
Example 11. Kolmogorov-Smirnov Test……210
Example 12. Kruskal-Wallis One-Way ANOVA……..215
Example 13. Wilcoxon Signed Ranks Test ……..220
Example 14. Friedman Two-Way ANOVA……225
Example 15. One-Factor, Between-Subjects ANOVA ……..312
Example 16. Two-Factor, Between-Subjects ANOVA ……..330
Example 17. Planned Comparisons……353
Example 18. Unplanned Comparisons……..362
Example 19. Analysis of Interactions……..377
Example 20. One-Factor, Within-Subjects ANOVA……..405
Example 21. Two-Factor, Within-Subjects ANOVA……..413
Example 22. Geisser-Greenhouse and Huynh-Feldt Corrections……..421
Example 23. Testing Order Effects in Balanced Latin Squares……..432
Example 24. Within-Subjects and Between-Subjects Design Comparison……..439
Example 25. Two-Way, Mixed-Factors ANOVA……..448
Example 26. Complete Hierarchical Between-Subjects Design……..504
Example 27. Partial Hierarchical Between-Subjects Design……509
Example 28. Simple Blocking of $2^k$ Within-Subjects Design ……546
Example 29. Complex Blocking of $2^k$ Within-Subjects Design……..550
Example 30. One-Half Replicate of $2^4$ Between-Subjects Design……557
Example 31. 4x4 Latin Square Designs……605
Example 32. Linear Correlation Coefficient ……..623
Example 33. Alternative Linear Correlations……636
Example 34. Simple Linear Regression……659
Example 35. One-Way Analysis of Covariance (ANCOVA)……..673
Example 36. Multiple Linear Regression……..732
Example 37. Best Regression Equation……..745
Example 38. Polynomial Regression……756
Example 39. Orthogonal, Between-Subjects, Central-Composite Design……..796
Example 40. Blocked, Within-Subjects, Central-Composite Design ……802

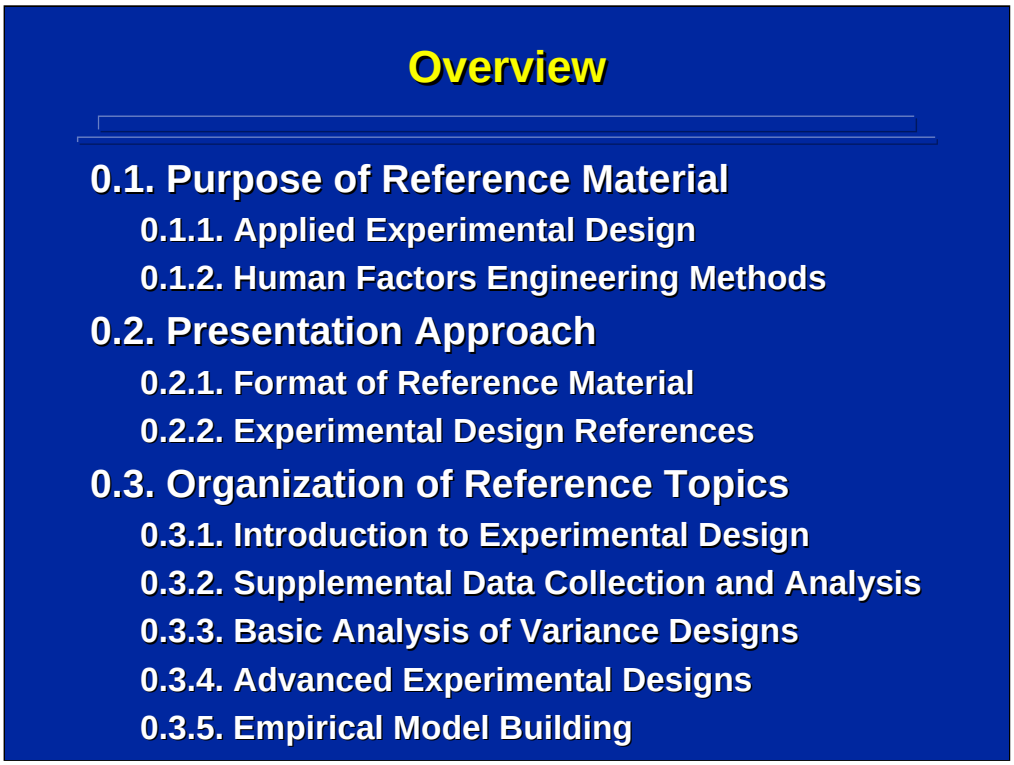
# Human Factors Experimental Design and Analysis Reference

**Dr. Robert C. Williges**
**Ralph H. Bogle Professor Emeritus**
**Grado Department of Industrial and Systems Engineering**
**Virginia Polytechnic Institute and State University**

**December 2006**

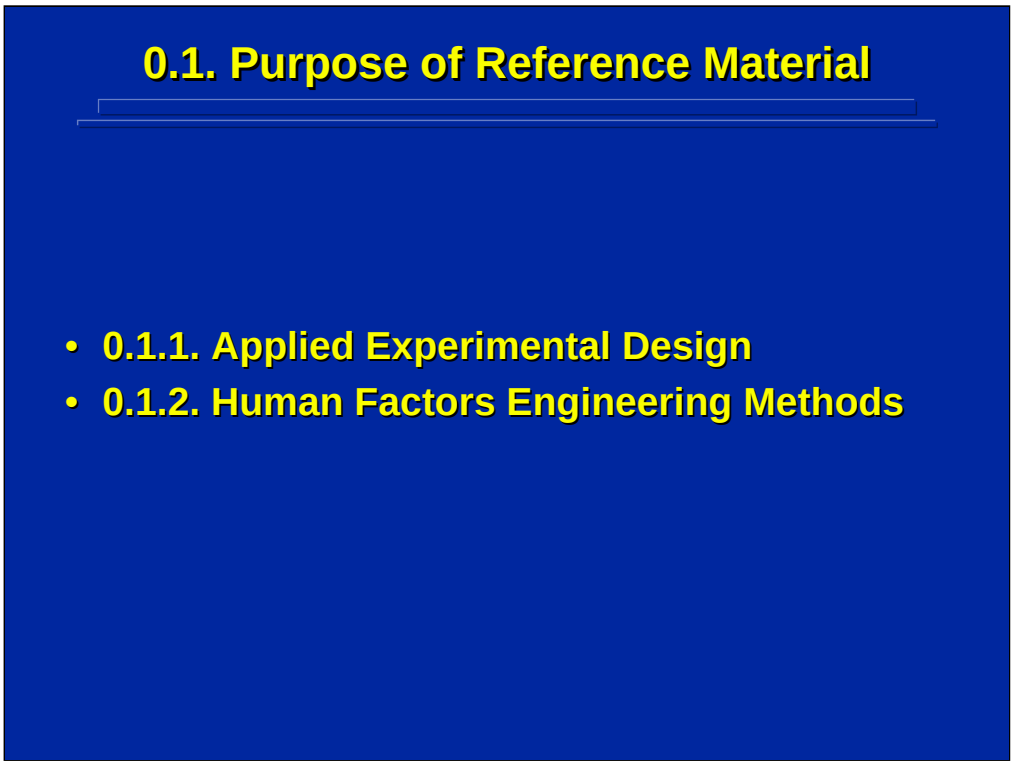
This set of slides on Human Factors Experimental Design and Analysis is designed to provide reference material to human factors engineers on research design and analysis techniques. This material is organized around the concept of a researcher's handbook that is available on a desktop computer and can provide an overview of critical experimental design concepts and methods for the human factors engineer as well as provide key references to the scientific literature related to these techniques.

It is assumed that users of this material are researchers in human factors engineering and ergonomics who have background in statistics and experimental design. These slides and accompanying notes provide reference material to help the researcher choose the appropriate experimental design and analysis. This reference material is not designed as a simple look-up for statistical procedures. Rather, it is designed to provide an overview and roadmap to techniques with reference to the statistical literature that provides details of procedures that should be reviewed before using them.

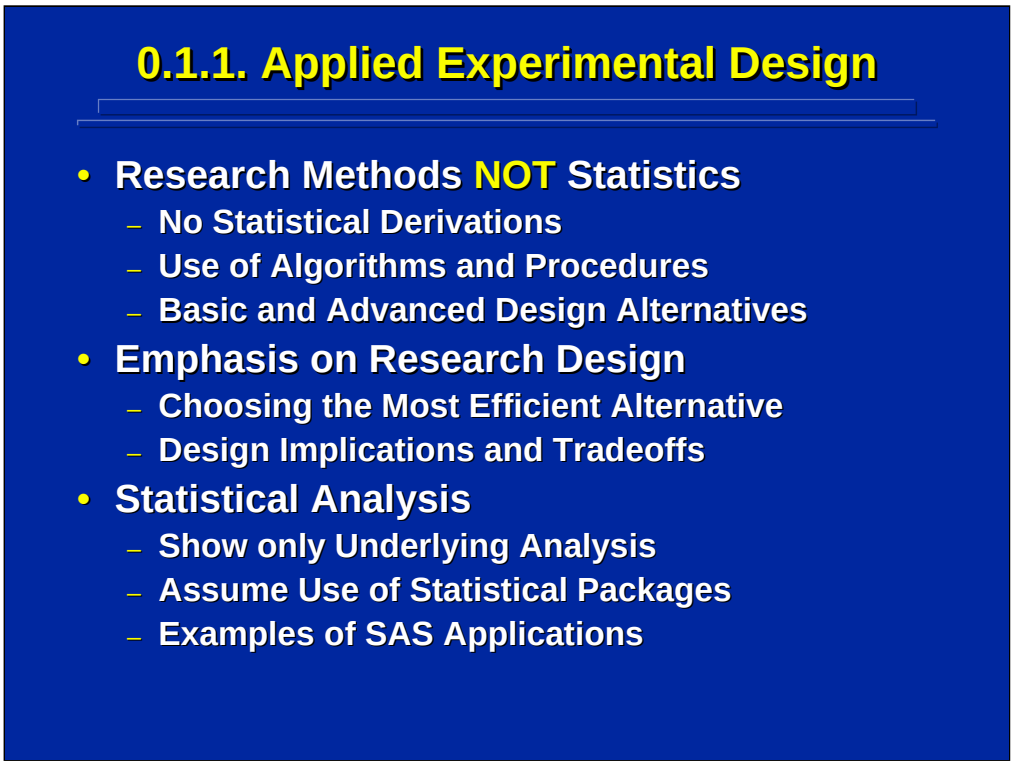



This is the outline of topics covered in the Overview to the reference material on applied experimental design. The purpose, presentation style, and organization of the topics are discussed in this overview.

Each of the subsequent major topic presentations in this reference material begins with an numbered outline of the subtopics covered. The detailed information content for every major topic follows this numbering system to facilitate user reference.



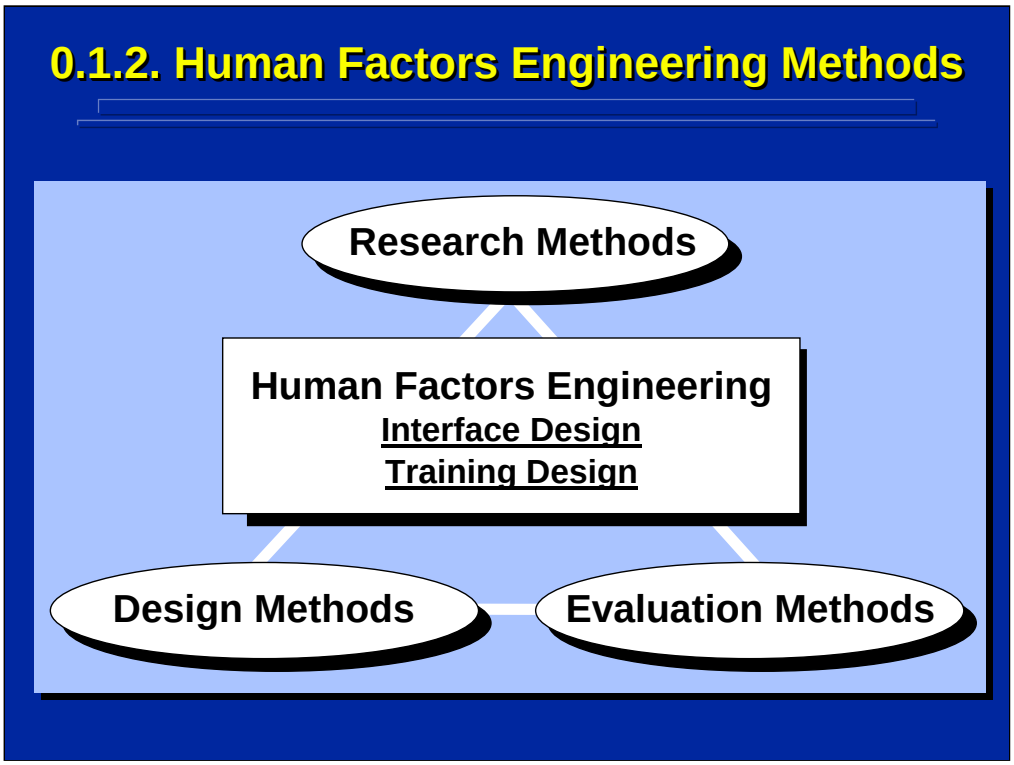

This is an example of the outline slide that introduces each topic subsection. As this outline suggests, the purpose of this reference material is to provide an overview of various applied experimental design procedures that are useful in human factors engineering. The implications of applied experimental design and its relationship to human factors methods are described in this subsection.



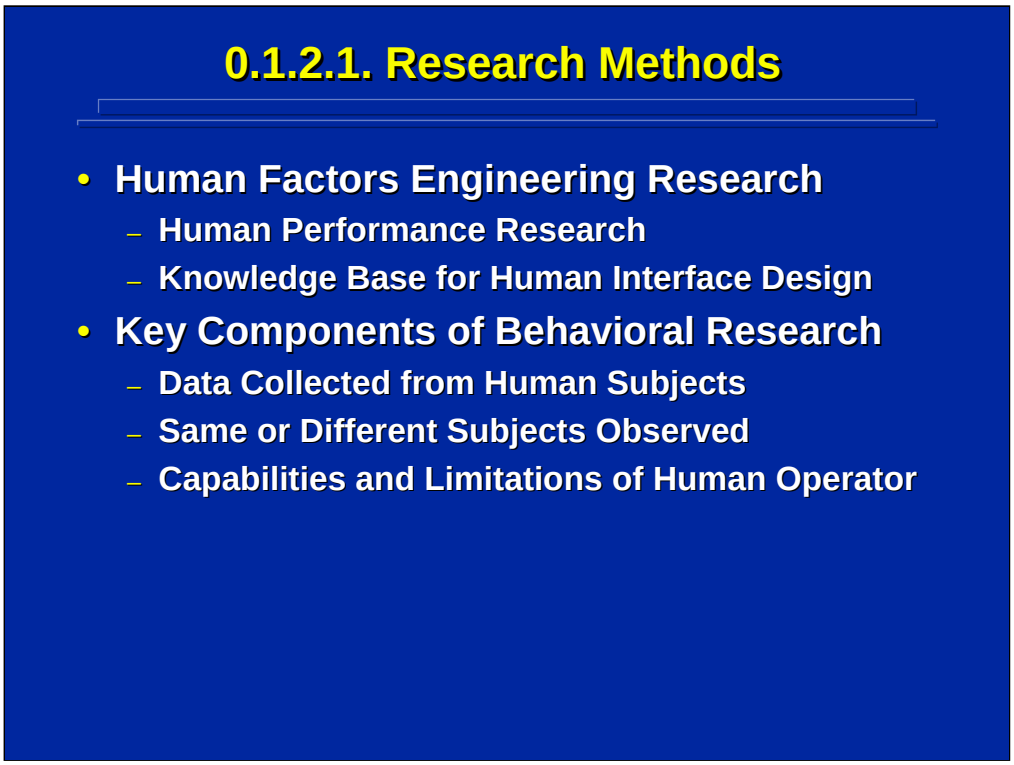

The emphasis of this reference material is on applied experimental design research methods and not on mathematical statistics. In lieu of statistical derivations, procedural steps and algorithms are presented for various experimental design calculations and representations such as statistical models, expected mean square, computational formulae, etc.

The reference material is designed to aid the human factors researcher in choosing the most efficient experimental design among a variety of available alternatives. Consequently, the various alternatives are outlined, and the tradeoffs among these alternatives are presented.

Examples of statistical analyses are provided for only the major procedures. It is assumed that most researchers will use a statistical analysis package to analyze their data. Consequently, most analyses shown in the reference material are presented in an appendix report by Slater and Williges (2006) that provides the program statements and output pages from the SAS application package.

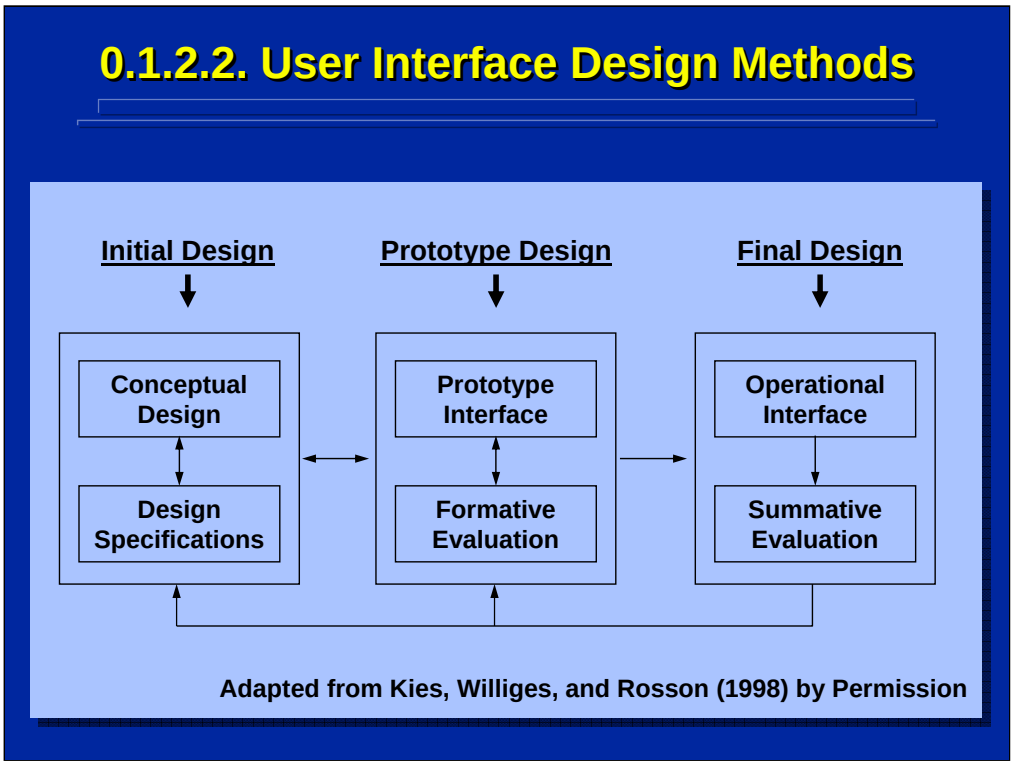



Human factors engineering involves both the human interface design of complex systems and the complimentary training of users of those systems. Successful human interface and training design requires understanding and mastery of various research, design, and evaluation methods. Applied experimental design is useful in each of these three major categories of methods.



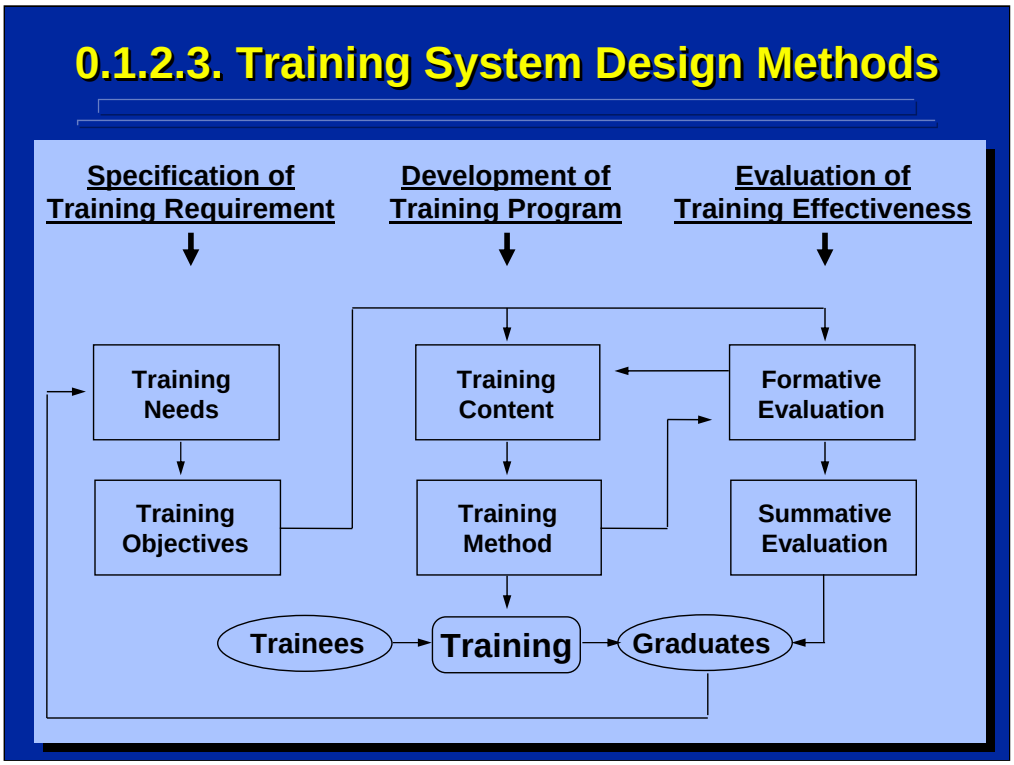

Experimental designs are central to human factors engineering research. This research deals primarily with human performance research that focuses on cognitive, motor, and biomechanical aspects of the human. Human factors engineering research provides the scientific knowledge base for human interface design in complex systems, and this research is based largely on experimental designs.

Human factors engineering research is characterized by three key components. First, the data are related to aspects of human performance and are collected from human subjects. Second, either the same sample of subjects is observed in a variety of treatment combinations or an independent sample of subjects is observed in each treatment combination. And, third, the research is focused on developing a scientific database of human operator capabilities and limitations in complex systems.

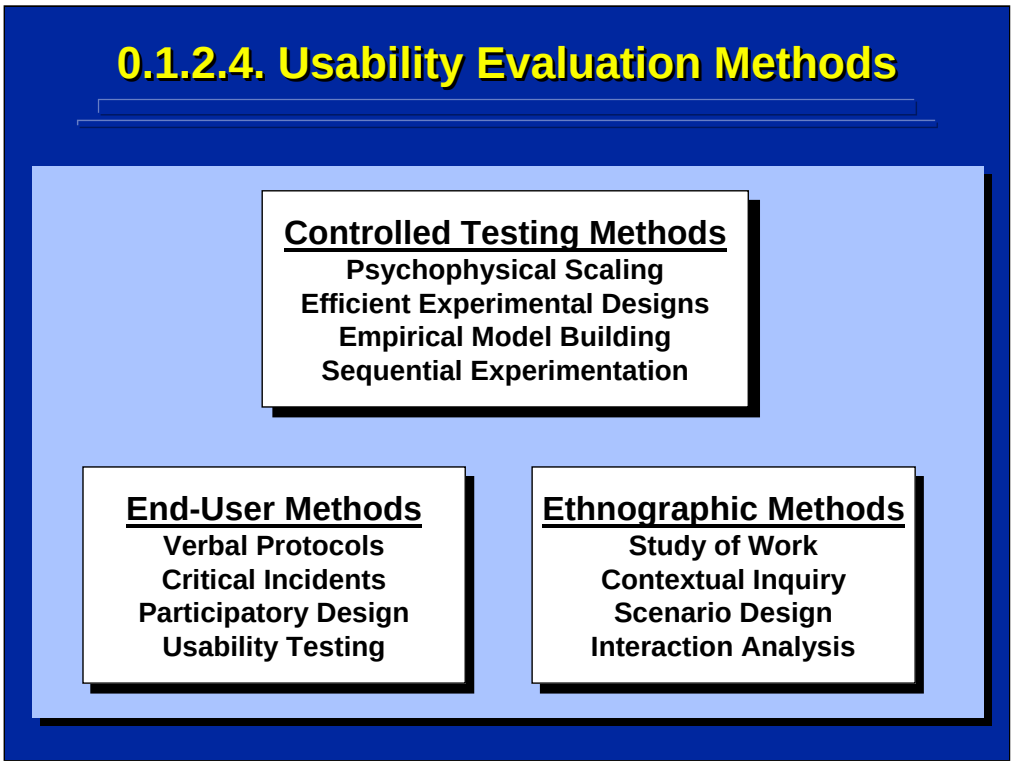



User-centered interface design is an iterative design process that is focused on the user of the system as shown by the two-headed arrows and the feedback loop in this figure that was modified from Kies, Williges, and Rosson (1998). In their article, they discuss appropriate ethnographic and experimental design methods for iterative design in each of three major phases of design of computer-supported cooperative work systems.

A variety of methods have been developed to support this design process. Essentially, these methods deal first with initial interface design to provide a conceptual design and specific design specifications. Next a prototype design of the interface is developed and actual users are tested somewhat informally though formative evaluation procedures in an iterative fashion. Following successful prototype design, the final operational interface design is developed and tested through a final, summative evaluation. Additional design iterations and major design revisions could be conducted as shown by the feedback loops in the figure. Rigorous experimental design procedures are most often used during summative evaluation in the user-centered design process.



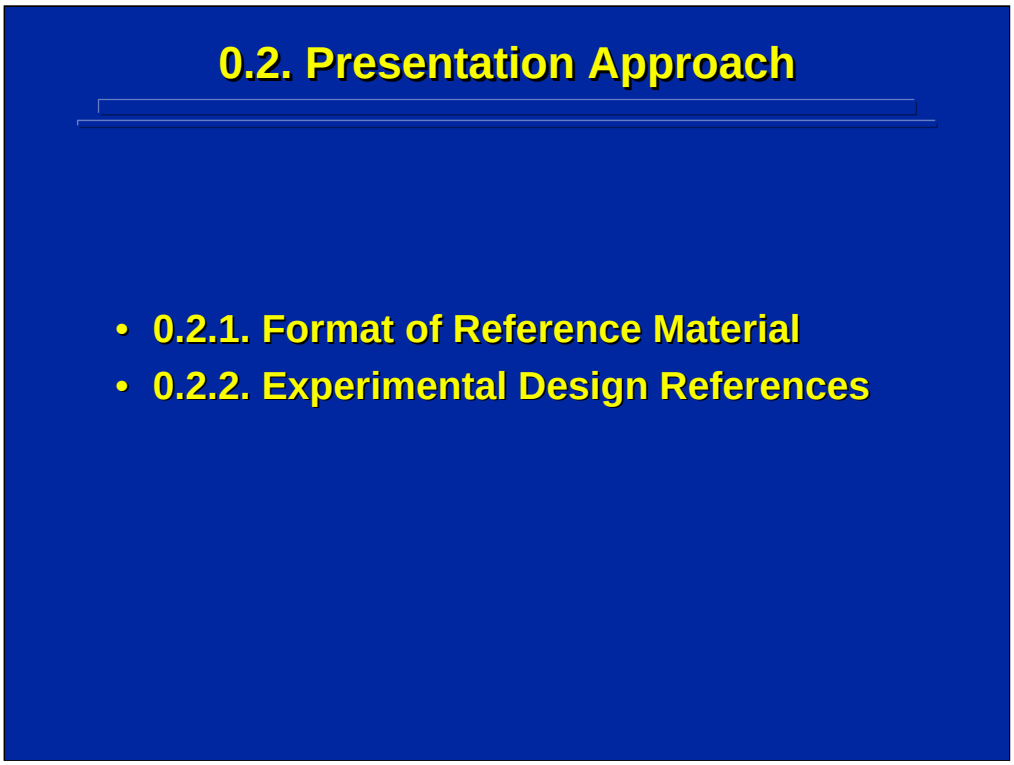

Most complex systems require human operator training in order to achieve the best system performance. The design of these training systems is also an iterative process that involves user testing. As shown in this figure, the three major stages of training system design include specification of training requirements, development of the training program, and evaluation of training effectiveness as discussed by Goldstein and Ford (2002). Experimental designs are used primarily in the summative evaluations of graduates of the resulting training system in order to evaluate the efficacy of training.



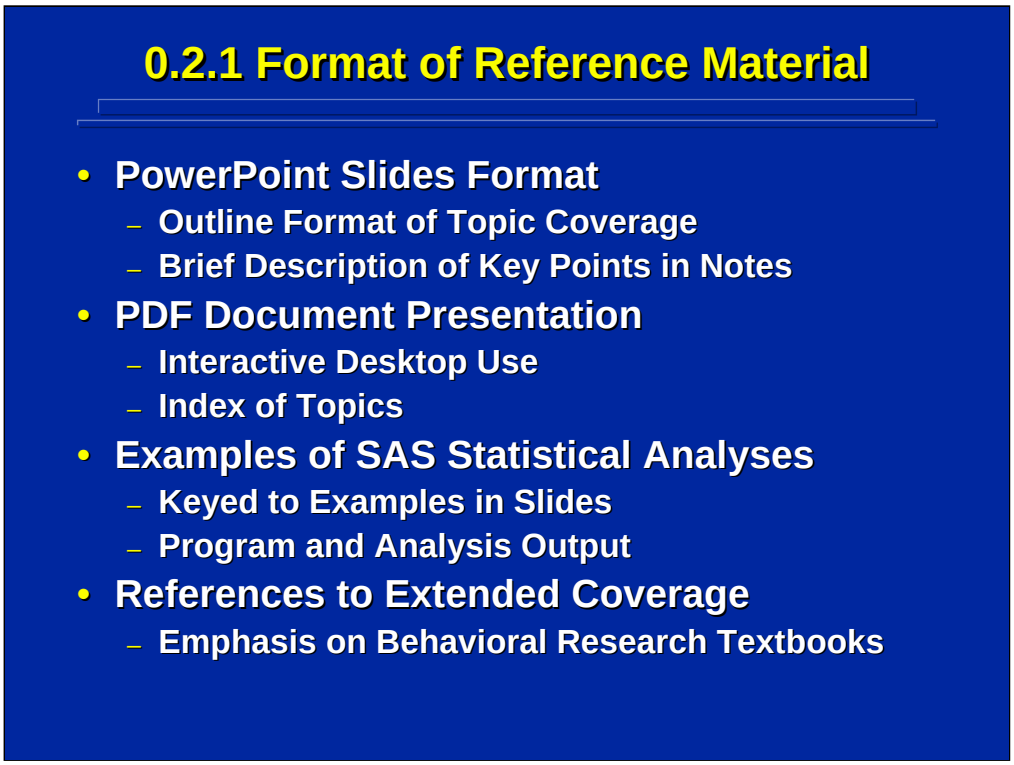

A variety of methods are available to support human factors engineering evaluation activities. As shown on this slide, these methods can be grouped into end-user, ethnographic, and controlled testing methods.

End-user methods involve the user of the system in the evaluation process. Verbal protocols and critical incidents are discussed in more detail as techniques to support supplemental data in experimental design. Participatory design involves end-user participation and evaluation throughout the design process (Schuler and Namioka, 1993). Usability testing methods are focused specifically on issues related to improving user performance of the system primarily during formative evaluation in the iterative design process. Hartson, Andre, and Williges (2003) provide a detailed breakdown of usability testing methods into expert, user, model, and location of usability evaluation methods across a variety of criteria.

Kies, Williges, and Rosson (1998) discuss appropriate ethnographic and experimental design methods for formative and summative evaluation of socio-technical systems. Experimental designs are examples of controlled testing methods. Efficient experimental designs, empirical model building, and sequential experimentation are most useful in complex system research and design.



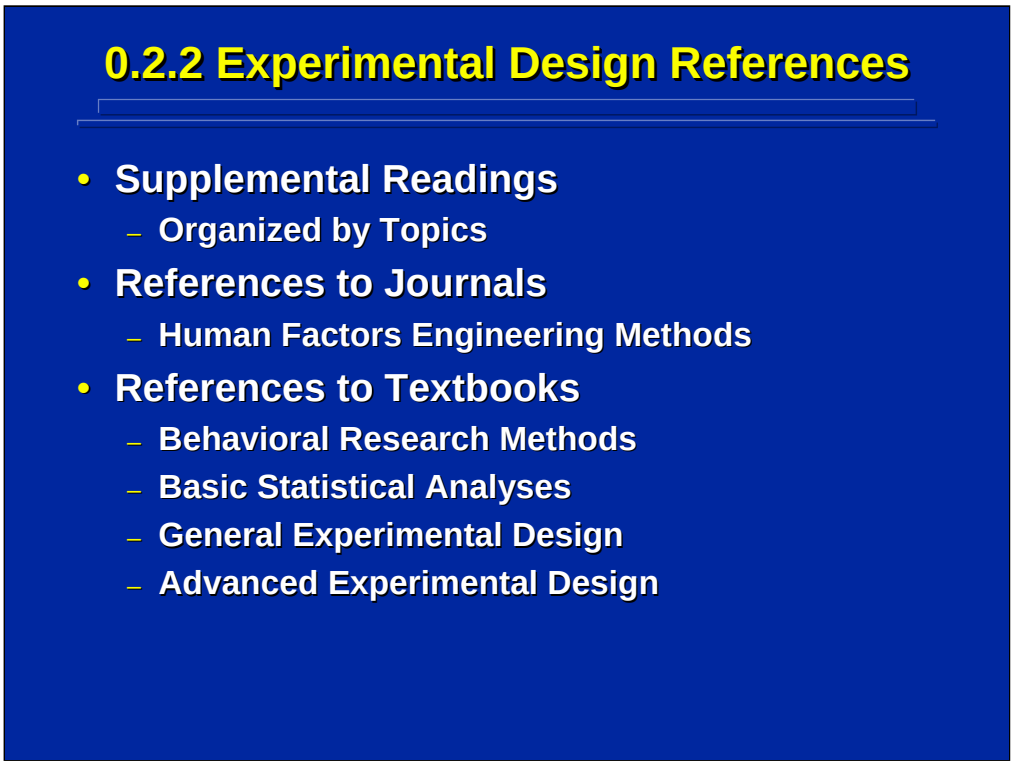

The presentation used in this reference material is focused on a researcher's handbook approach. Both the format of the material and the scientific references are directed toward material to support the human factors engineer who is planning, conducting, analyzing, and reporting results of experiments.



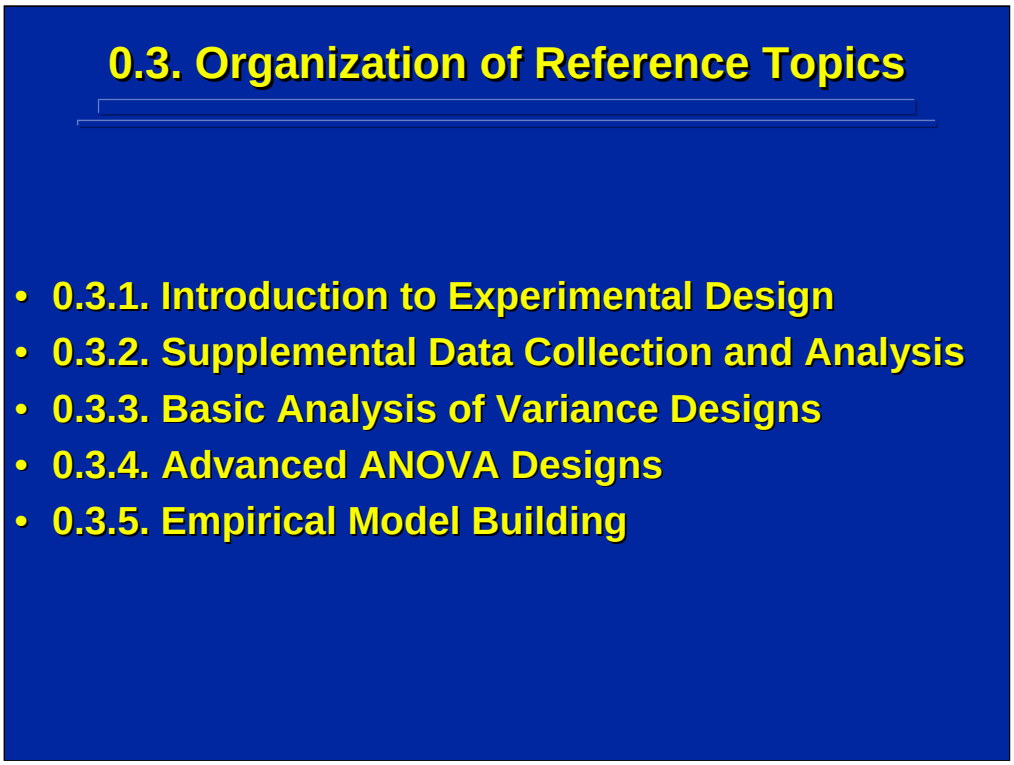

The reference material was prepared in a PowerPoint slide format. Each page of the reference shows a slide with the material presented in an outline format. Notes are provided under each slide to provide a brief description of the outline and to emphasize the major points of each slide. All of the reference material is delivered in PDF format to facilitate cross-platform, desktop computer use by the human factors engineer. Bookmarks are provided to a subject index in the PDF file.

Throughout the reference material, formulae are presented for statistical computations and examples are provided for the major computations. Additionally, these examples were calculated on a statistical package using SAS as an example. The data inputs, procedures statements, and computational outputs of SAS are provided in an appendix (Slater and Williges 2006) that is hyper-linked to the reference slides so that researchers can view detailed examples of using a statistical package for computations. Each example in the Slater and Williges (2006) appendix is also linked directly to the SAS editor.

References to the scientific literature are provided throughout. References are also provided for behavioral science textbooks on experimental design that can be used for supplemental reading on a more detailed coverage of each topic covered. The complete citation for each reference is listed at the end of the PDF file.

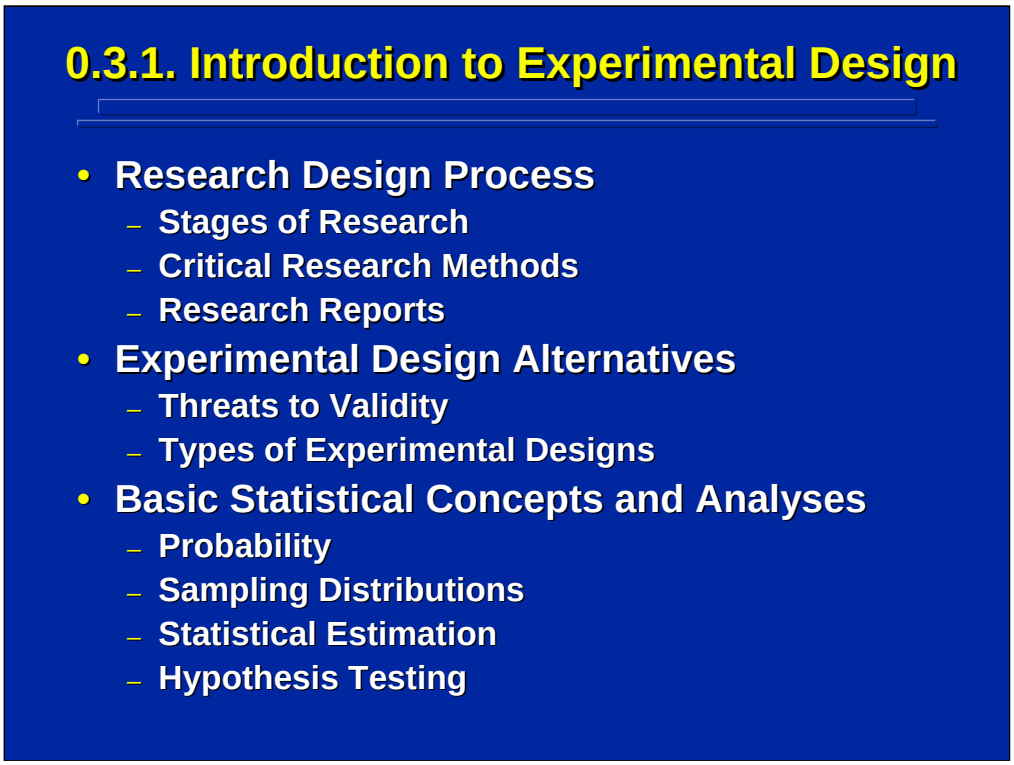



Each topic in the reference material lists supplemental readings. These supplemental readings provide more detailed coverage for a better understanding of each topic. Reference to key methodological articles in the human factors journals and textbooks are provided. In addition, references on behavioral research methods, basic statistical analyses, and experimental design textbooks are provided as appropriate.

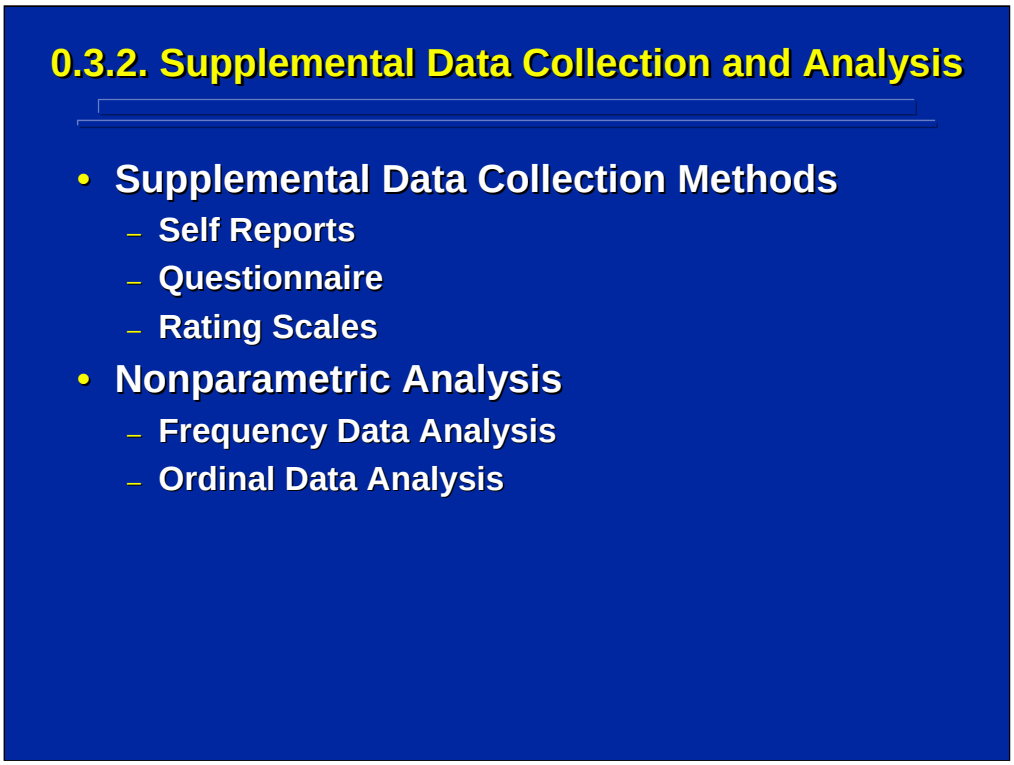



The reference material is organized around five major sections. These sections cover general considerations in experimental design, supplemental data collection and analysis, basic analysis of variance experimental design, advanced experimental design, and empirical model building.

Section 1 covers topics related to critical aspects of the experimental design process used by human factors engineers. Section 2 covers methods of data collection and analysis of supplemental data that are often collected in addition to the major data collected through experimental designs. Section 3 addresses concepts of basic analysis of variance (ANOVA) designs used by human factors researchers for collecting data on human subjects performing tasks in complex systems environments. Section 4 covers advanced experimental design topics that are useful to human factors engineers who must deal with procedural constraints in data collection and large-scale data collection efforts. Finally, Section 5 describes empirical model building procedures used to predict human performance in complex systems.




The introduction section to experimental design covers three major topics. These topics include the research design process used by the human factors engineer, experimental design alternatives (i.e., quasi-, and randomized experimental designs), and basic statistical concepts and analyses needed for experimental design. These three topics are covered in Section 1 of the reference material.

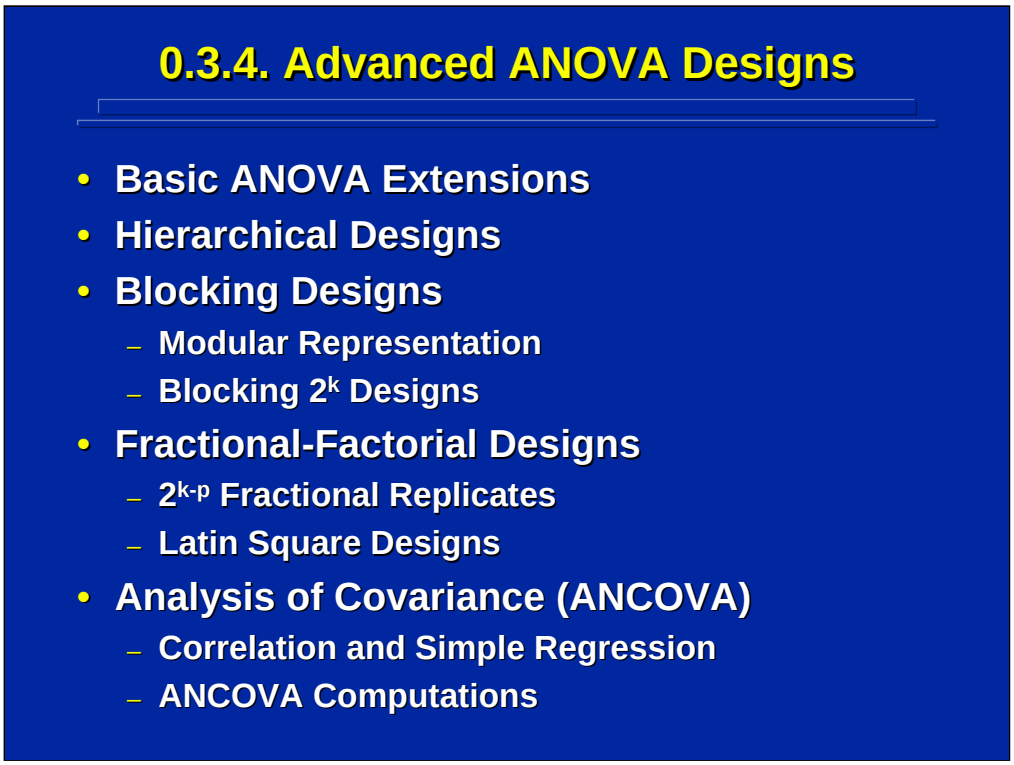



Supplemental data collection and analysis involves additional data collected on human subjects to aid in the understanding of the results obtained from the experimental design. Two topics are covered in this section. First, an overview of supplemental data collection methods is discussed with an emphasis on rating scales. Second, a summary of the most common data analysis procedures for supplemental data consisting of frequencies and rank orders are presented. Both of these topics are also covered in Section 2 of the reference material.

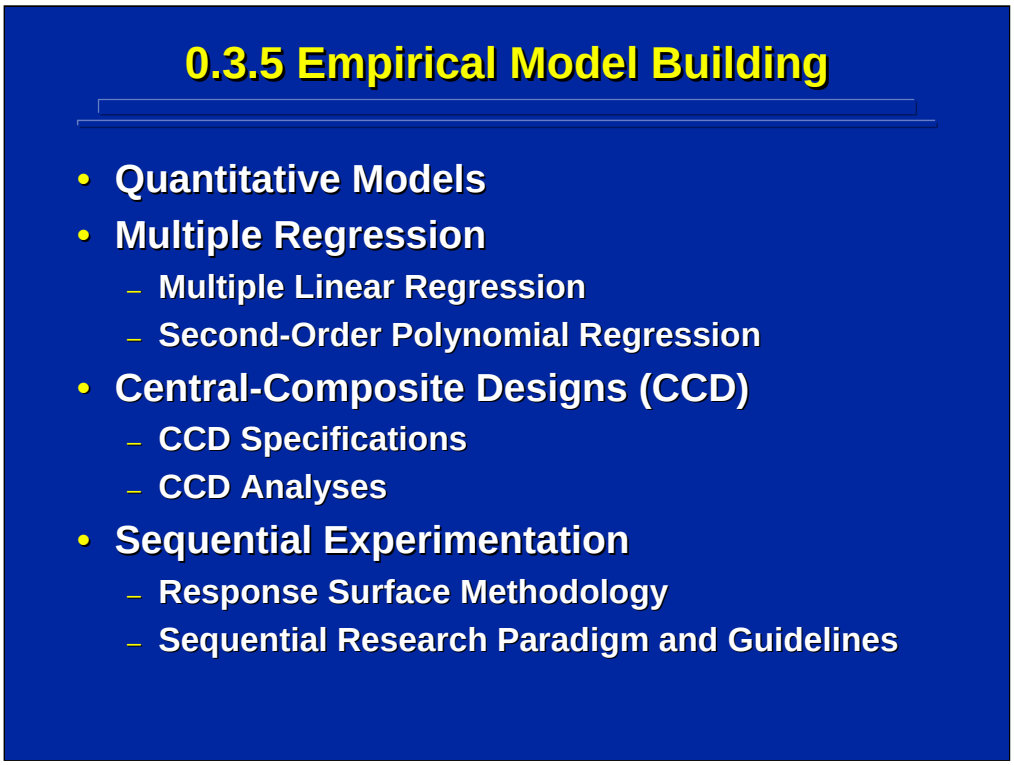
## 0.3.3 Basic Analysis of Variance Designs

- Analysis of Variance (ANOVA) Classification
  - Basic Terms
  - Design Alternatives
  - ANOVA Summary Table Components
- Between-Subjects Design
  - One, Two-, and n-Factor Designs
- Analysis of Comparisons and Interactions
  - Paired-Comparisons
  - Evaluating Interactions
- Within-Subjects Design
- Mixed-Factors Design

Factorial analysis of variance designs are the major experimental designs used by human factors engineers. The reference section on basic ANOVA covers five major topics including analysis of variance design classification, between-subjects or completely randomized designs in which a different group of subjects is used in each treatment condition, post hoc analysis of paired comparisons and interactions, within-subjects or repeated measures designs in which the same subject is used in all treatment conditions, and mixed-factors or split-plot designs in which some treatment conditions are between-subjects conditions and some are within-subjects conditions. Each of these topics is covered in Section 3 of the reference material.

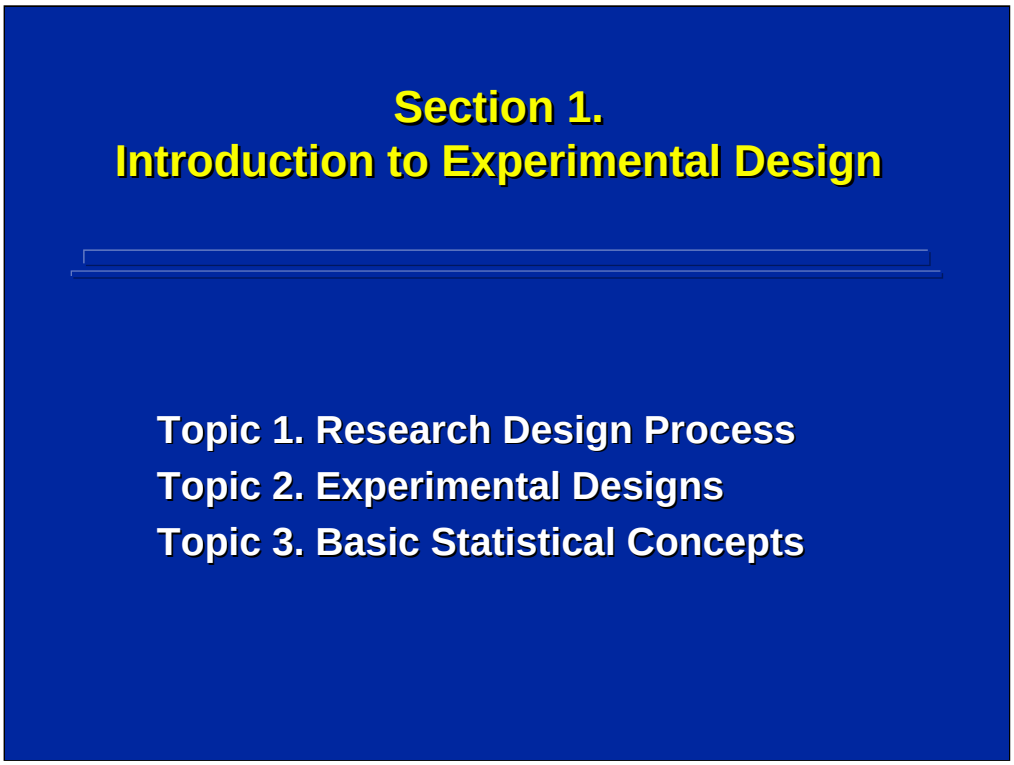



This section of the reference material covers major advanced experimental design and analysis procedures used by human factors engineers to handle certain experimental constraints encountered in research. These advanced designs are built on basic ANOVA and regression analysis. Topics covered in the advanced experimental design section include extensions of basic ANOVA, hierarchical or nested designs, blocking designs, fractional-factorial designs, and fundamentals of simple regression analysis used in the analysis of covariance. These five topics are covered in Section 4 of the reference material.




The final section of the reference material covers empirical model building procedures. Four major topics are covered. The section begins with a discussion of quantitative models in research that are used to predict human performance. Next empirical model building using polynomial regression with central-composite designs are described. Finally, sequential experimentation that involves a series of small related experiments covering an extremely large data space are described as a paradigm for conducting systematic research on complex human factors problems. All of these topics are covered in Section 5 of the reference material.

Due to the building block approach used in presenting the topics covered in this reference material, some questions raised in earlier sections are answered in later sections. The user should user the interactive aspects of this reference to locate expanded discussion of some topics throughout the presentation.




By way of introduction, Section 1 summarizes some major components that are fundamental to experimental design and analysis. This section covers:

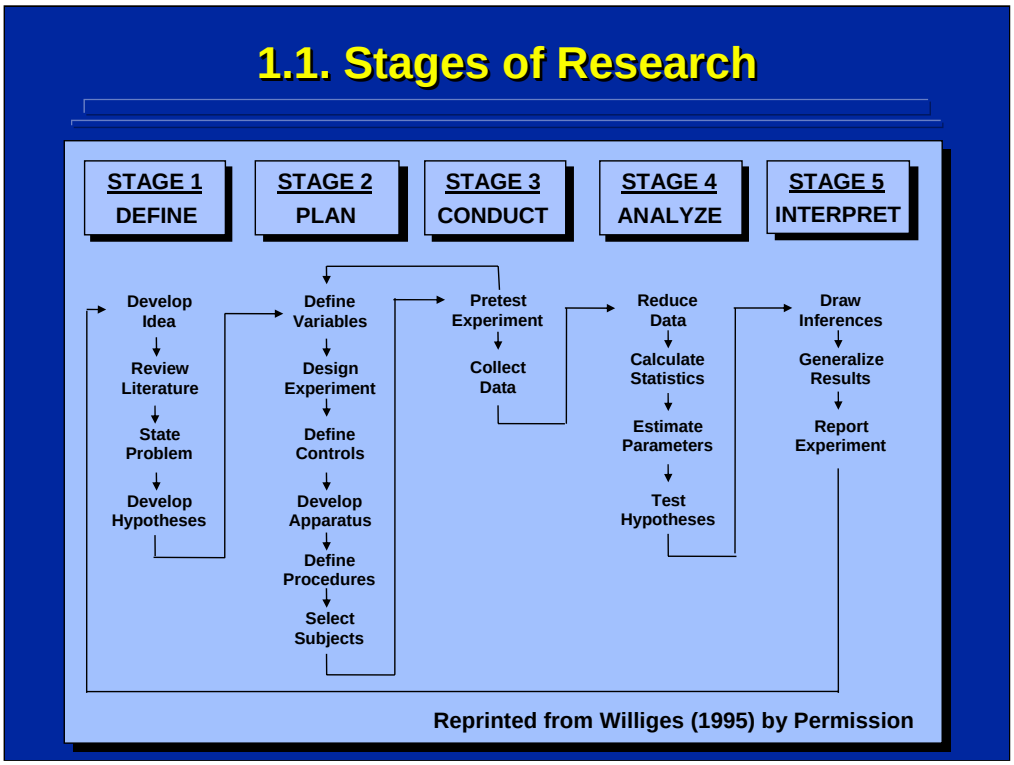
Topic 1 - the research design process;

Topic 2 - major categories of experimental design alternatives; and

Topic 3 - a brief review of basic statistical concepts and analyses used in experimental design.

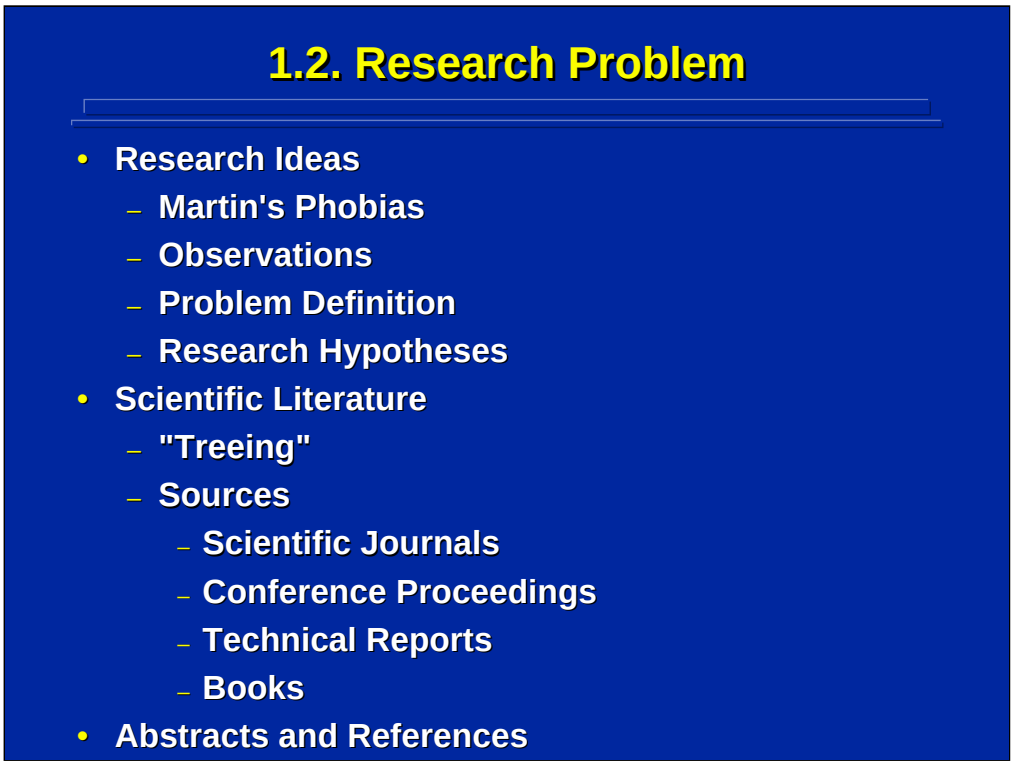
This topic is an introduction to experimental design that deals with the overall research design process. First, the various stages of research are presented in a flow diagram. Next six critical aspects of this process are highlighted beginning with the Research Problem through Research Reports.

As with all subsequent topics covered in the reference material, this topic concludes with a summary followed by suggestions for supplemental readings for in-depth coverage of the material covered in this topic. Due to space restrictions, the complete citation for each supplemental reading is not presented on the summary slide. However, the complete citation is presented in the References section that is bookmarked in the PDF file.




Williges (1995) presented a research process with five inter-related stages as depicted in this slide. (This figure is reprinted by permission of Person Education, Inc., Upper Saddle River, New Jersey.) His five stages include defining, planning, conducting, analyzing, and interpreting. Often, an experimenter only thinks of research design and analysis and fails to consider all five stages of the research process. Note that this process is a closed-loop flow of several considerations leading to successful research.

Several important research procedures related to the Williges (1995) five-stage research process are subsequently covered in this topic to highlight major issues that can cause problems in the research enterprise. These procedures begin with the definition stage of research.



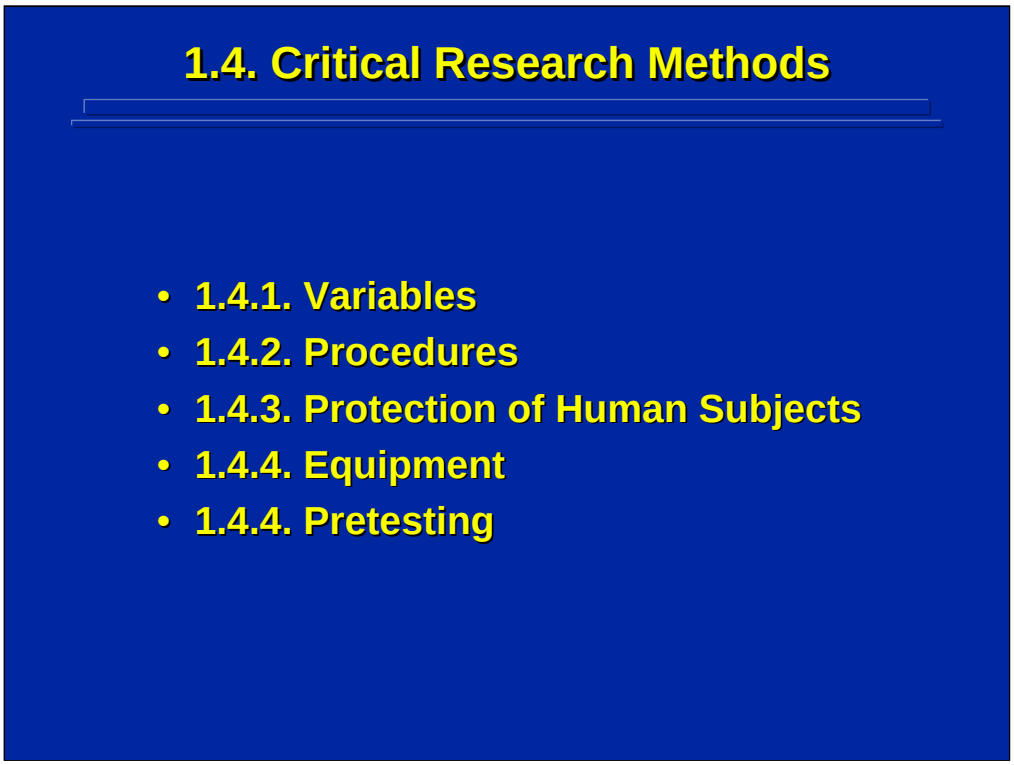

Martin (2004) humorously discusses many common apprehensions that new researchers have in conducting research, but remember that the possibility of exactly replicating existing research is quite remote. One should try to state the research problem in one paragraph, and then state the hypothesis to be tested through data collection.

An efficient way of searching the scientific literature is a technique called “treeing”. The researcher reads a recent article related to the research problem and then reviews the articles in its reference list. Always be sure that you read any reference that you cite to insure accuracy. Do not rely on secondary references. Online searches and electronic publishing can facilitate searching the scientific literature.

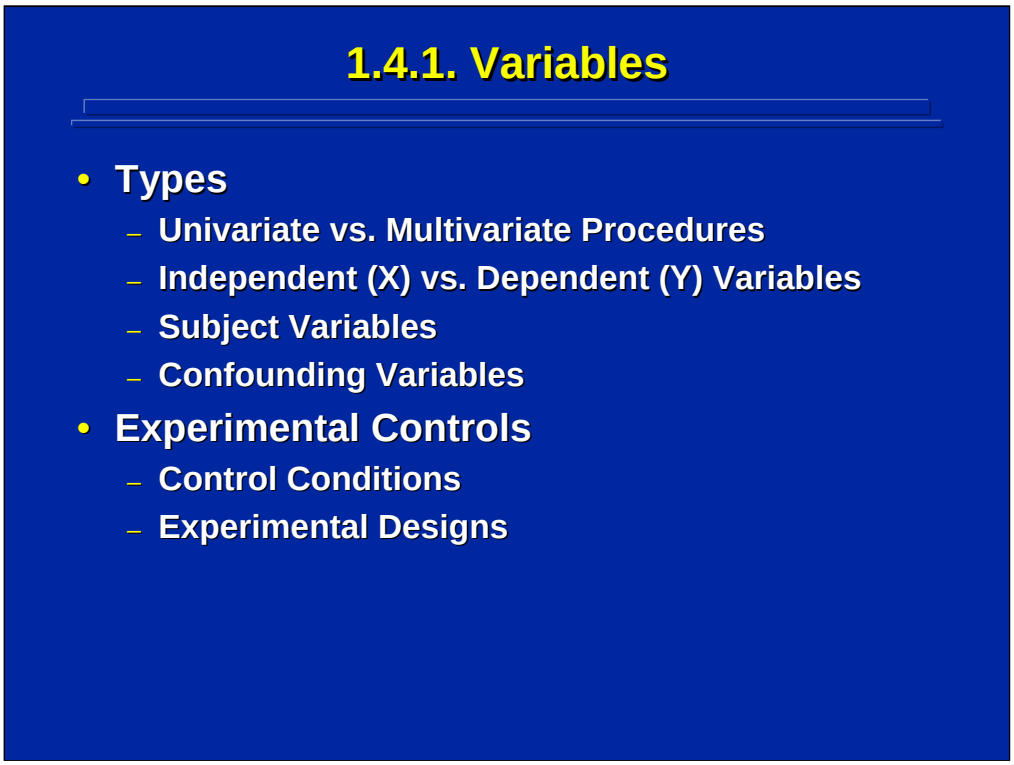
Two things to consider in reference sources are the scientific rigor and the age of the material. Scientific journals have an editorial review board to enhance rigor, but the review and publishing process may take years. Conference proceedings include the most recent research, but are often only reviewed on the basis of an abstract. Technical reports are reports published by individual laboratories usually without external review. Many books have review chapters that summarize older literature in a research area. A researcher should be compulsive and write notes or an abstract on each article read as well as the complete reference citation.

The scientific method uses experimental designs that require systematic observation during data collection. So, one defines the specific circumstances under which observations are made. Extraneous variables are controlled to avoid confounding effects and to facilitate interpretation. Human behavior is observed in an unbiased, objective fashion that avoids experimenter opinions. The emphasis is placed on collecting quantified data so that inferential statistical analysis can be conducted on the resulting data set. From these results, one can infer causative relationships.

Besides insuring that the observations are repeatable, objective, and quantitative, the researcher should include as many relevant variables as possible in the investigation so that the results will generalize to real world applications. When all possible variables are operating, there is less control and more random error is added to the experiment. When designing an experiment, one must trade off which variables are controlled and which variables are not controlled to facilitate generalization. This often results in including several variables in one experiment and increases the data collection effort.

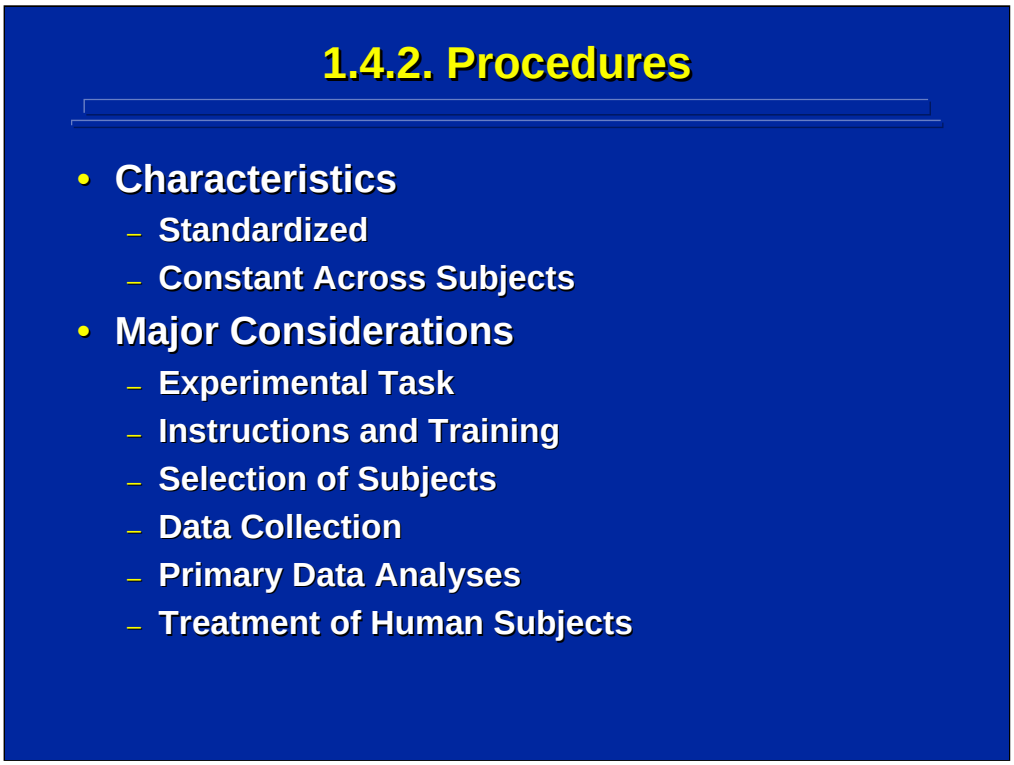



Research methods include topics such as variables, procedures, protecting human subjects, equipment, and pretesting. These four topics are critical because each can often result in major problems in the research process. Each is reviewed separately. Martin (2004) provides a more comprehensive discussion of these topics as well as other methods to consider in designing human subject research.



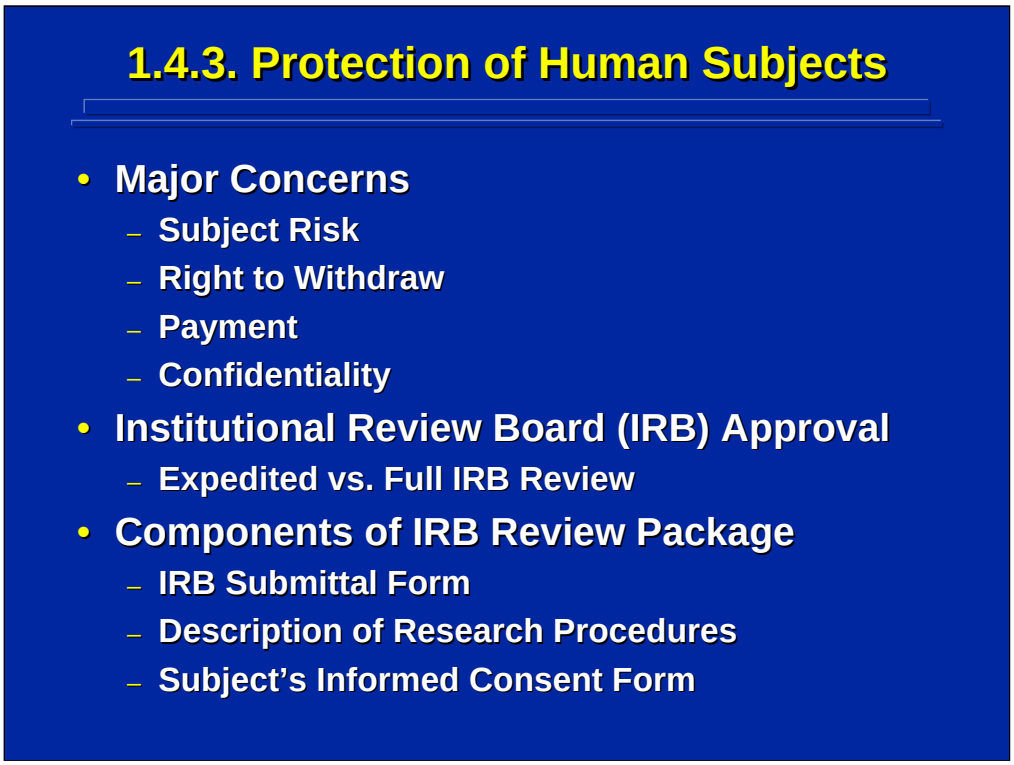

There are several types of variables used in discussing experimental designs. Univariate means consideration of one variable while multivariate means consideration of more than one variable. An independent variable (X variable) is a variable that the experimenter manipulates and is independent of the performance of subjects participating in the experiment. A dependent variable (Y variable) is one that depends upon the performance of the subjects in the experiment and constitutes the data collected in the experiment (e.g., errors, completion time, or accuracy). Subject variables are things such as prior experience that one tries to control through randomization or selection. Confounding variables are other variables that occur in the experiment that can affect the experiment but have nothing to do with the focus of the study.

Specific experimental designs are often chosen to control confounding variables. In most human factors research studies, one conducts multivariable experiments involving several independent variables simultaneously. However, human factors researchers usually conduct univariate statistical analyses on each dependent variable separately. Consequently, univariate data analyses rather than multivariate analyses are emphasized in this reference material.



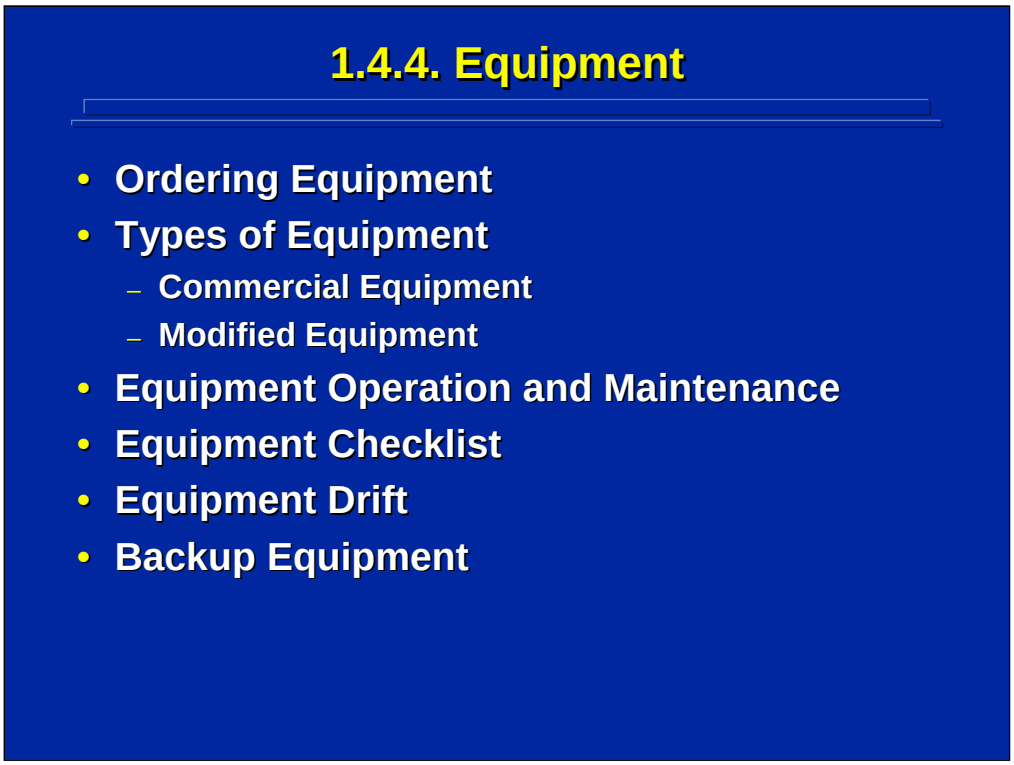

The keys to setting up procedures in an experiment are standardization and consistency in procedures across subjects. The task to be completed should be the same for each subject. Instructions and training should be written out and recorded for each subject so that everyone gets the same information. For example, recorded instructions should be played while the subjects are reading them so that they are forced to go from the beginning to the end at a constant rate. The selection of subjects should be representative of the subjects in the population of interest. Data collection should be systematically stored for accurate future reference. One should keep back ups for all data collection. And, the primary data analysis should be planned before data collection begins.

Treatment of human subjects is very important, because all human factors experiments are conducted using human subjects. Subjects should not be endangered physically or mentally during their participation. Since subjects are volunteers, they have the right to withdraw from the experiment at any time.



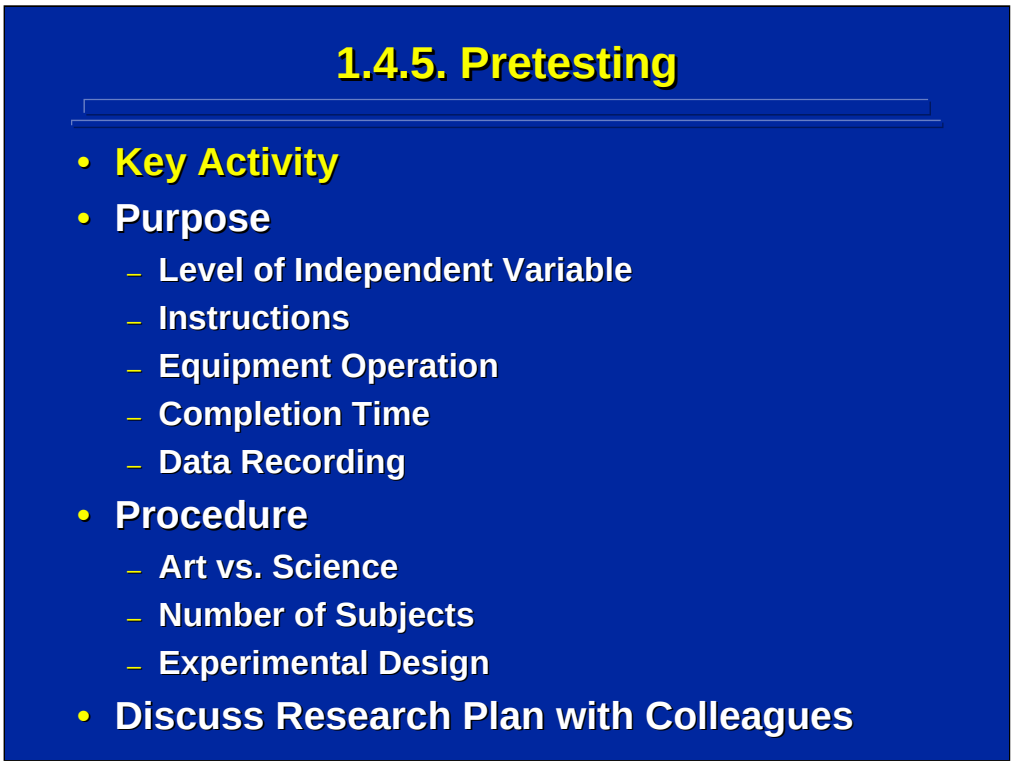

Major concerns in the protection of human subjects include subject risk, the right to withdraw, payment plans, and maintenance of confidentiality. For example, refer to subjects by number rather than name in data collection sheets to insure anonymity. Subjects in human factors experiments are often paid for their participation. If so, the researcher should be careful that payment does not interfere with the subject's right to withdraw.

Often an Institutional Review Board (IRB) assesses the level of subject risk during an experiment. If so, one must have IRB approval to proceed with the experiment. Two types of review are expedited and full IRB review. Most human factors research requires only expedited IRB review, because subjects are at low risk. If, however, minors are used as subjects or invasive procedures such as blood testing is involved in the human factors research, full IRB review is required. The standard IRB review components include an IRB submittal form, description of research procedures, and the subject's informed consent form.



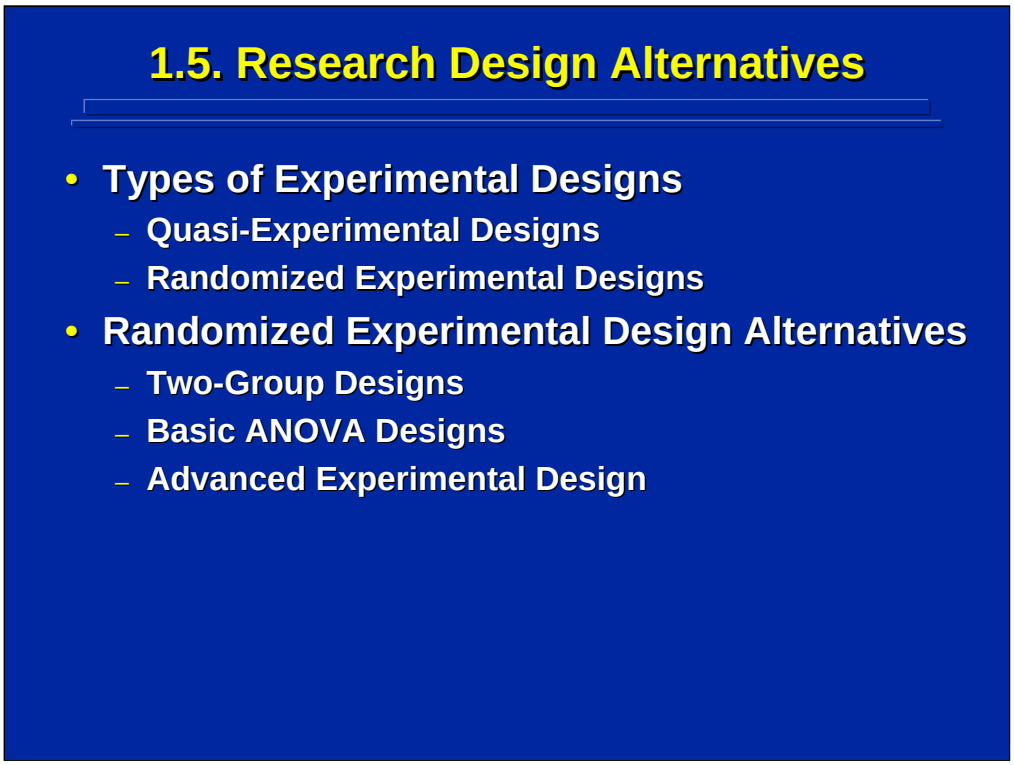

Sometimes equipment must be ordered and can delay the start of an experiment if ordering time is not considered. Both commercial and modified equipment is used in human factors research. The equipment must be set up the same way each time and must be maintained to avoid failure in the middle of data collection. An equipment checklist should be used to enforce consistency.

One must be careful of equipment drift where equipment settings or resolution could change over time as the equipment is used repeatedly. Analog equipment is more sensitive to equipment drift than digital equipment. So, sufficient warm-up period should be allowed for analog equipment before commencing data collection. If possible, the researcher should have backup equipment to avoid delays resulting from equipment failure.



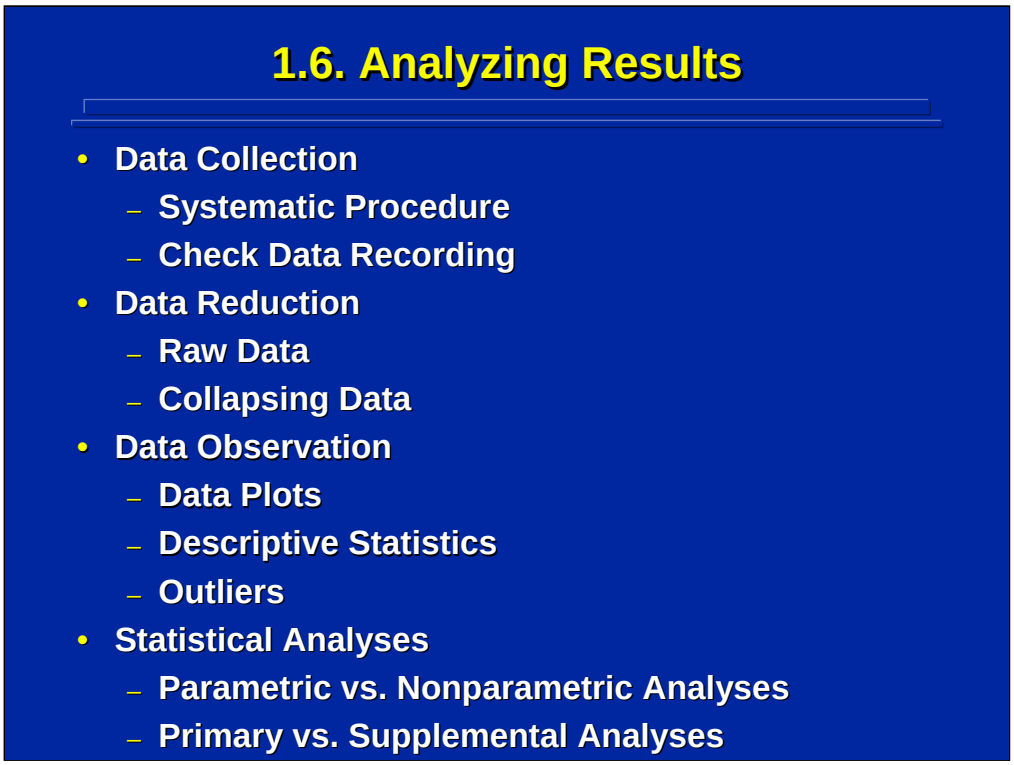

Pretesting is the most important aspect of setting up an experiment, but it is often overlooked or minimized. The purpose of pretesting is to check the levels of the independent variables to determine if they are appropriate. Instructions must be tested, because subjects may interpret instructions differently from the experimenter's intention. Equipment operation and completion time should also be pretested, because each subject will work at a different pace. Pretest data recording so that it is reliable and unbiased to insure that no data will be lost.

Pretesting is more of an art than a science. It takes experience and knowledge of the problem area. The pretesting procedure is really not set. The number of pretest subjects varies for each experiment. One subject is definitely not enough, so at least several subjects should be used. There are no formal experimental designs for pretesting. Usually one picks a treatment combination where a large difference is expected to check if these differences exist and if adjustments are needed. Finally, it is quite helpful to discuss plans with research colleagues who have experience with collecting data in a similar environment. They can provide good advice and insights on the pending experiment.




Experimental designs provide plans for the systematic collection of data under managed conditions as compared to making only passive observations. Cook and Campbell (1979) describe two general categories of experimental designs, quasi-experimental designs and randomized experimental designs. The distinction between them is determined by the existence of experimental control and random assignment of subjects. Quasi-experimental designs may or may not specify control conditions to manipulate in an experiment and do not have random assignment of subjects to treatment conditions. Randomized experimental designs have controls built into the design and also have random assignment of subjects to treatment conditions.

This reference material concentrates on randomized experimental designs, because they provide the most valid data for causative inferences and the most generalizable results. These experimental designs extend from two-group designs, to basic factorial ANOVA designs, to advanced experimental designs. Most human factors researchers use basic factorial experimental designs due to the nature of their research problems.

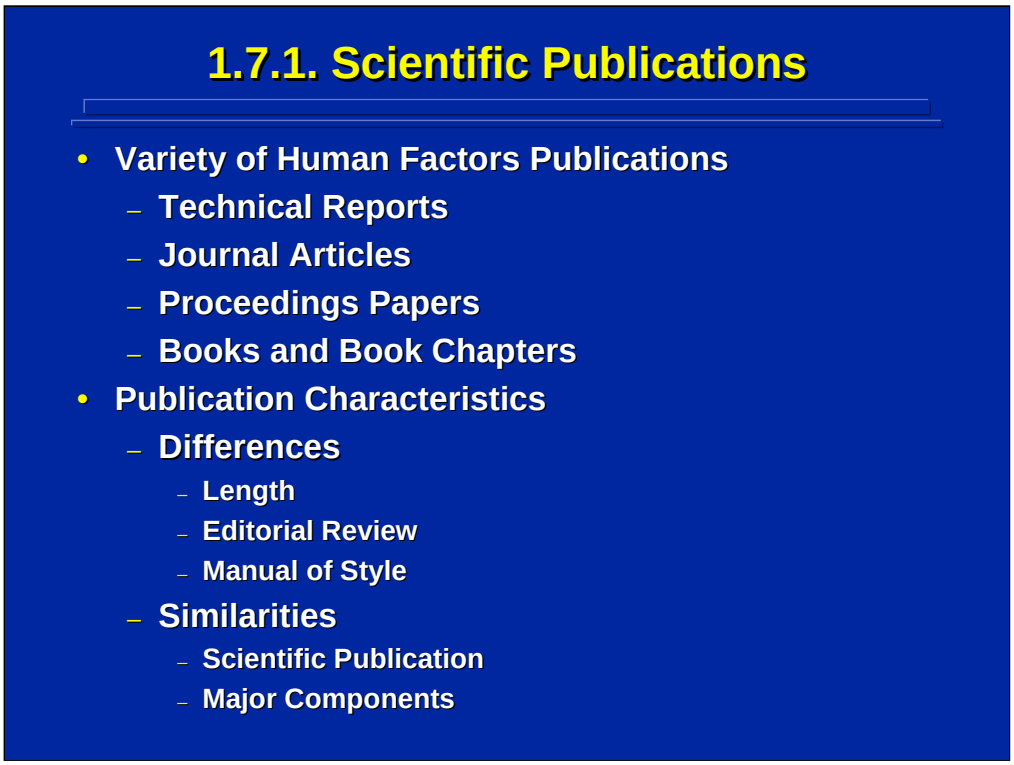



The experimenter should think about the major analyses before data collection to help in choosing the most appropriate design. Have some checks and balances built into data collection to insure accuracy. Usually some data reduction is necessary before conducting statistical analyses. Always keep your raw data at least until the report is written. One can always collapse data, but one cannot return to raw data after collapsing if secondary analyses should require using raw data.

Before conducting any statistical analysis, plot the data to determine if the expected differences seem to exist and the data are coded correctly. Looking at the results before analysis helps in interpreting the statistical analysis. Use descriptive statistic like means or variance in data plots. A good rule is never discard a data point unless one has clear documentation that it is an outlier and not a true reflection of subject variability.

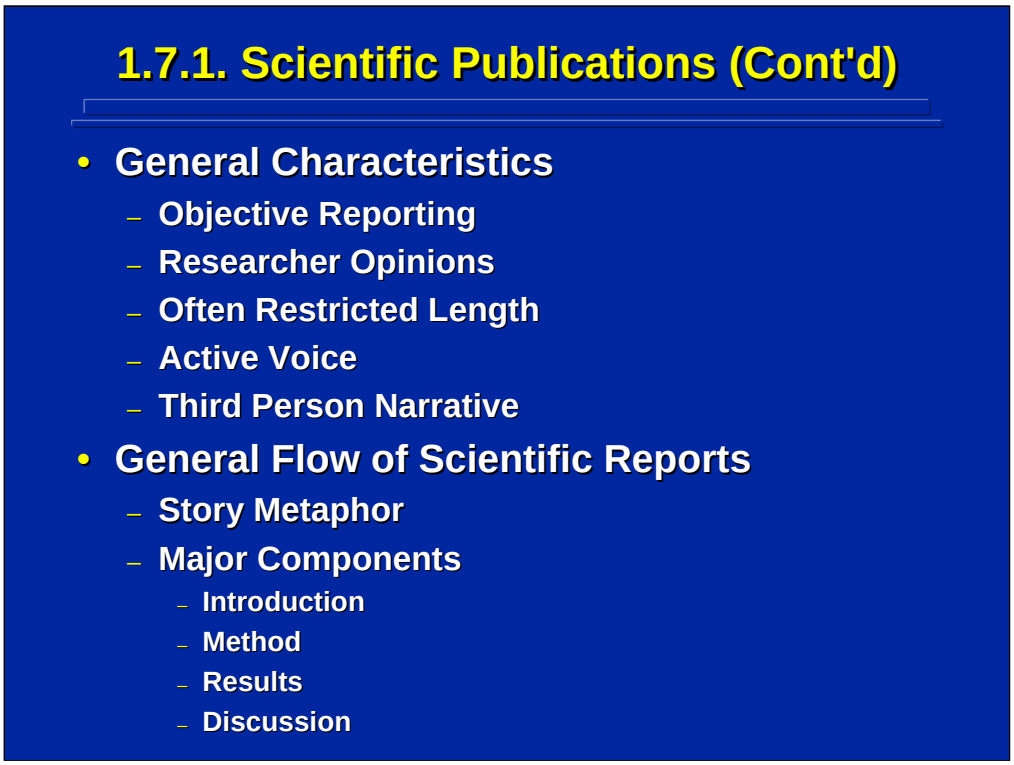
Parametric analyses have certain parameters that define the statistical analysis and have certain assumptions about the type and distribution of scores that are not needed in less powerful nonparametric analyses. Primary analyses are the major analyses that were planned before data were collected. Supplemental analyses aid in interpretation and are often based on nonparametric analysis of demographic data or ratings.

No piece of research is really complete until it is reported. Researchers have an obligation to their scientific colleagues to report their findings. A written report is most common, but reporting can also be an oral presentation. Several types of scientific publications are used, but each of them have major components in common while differing in other special sections and considerations.



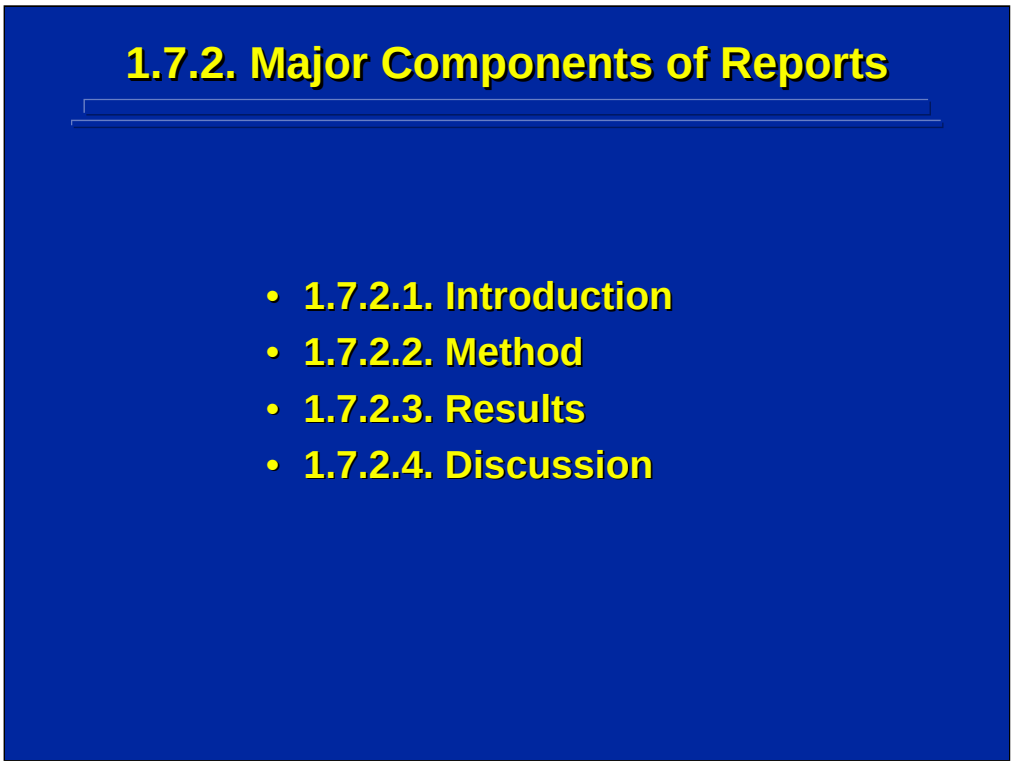

There are a variety of human factors publications. Technical reports are reports that are completed in the individual laboratory and submitted to research sponsors. Journal articles are publications that add to the archival scientific literature either in printed or electronic format. Proceedings papers are presented at scientific meetings like the HFES conference and are commonly published in CD-ROM format. Books and book chapters are part of the basic scientific literature.

Types of publications differ in length. Usually proceedings papers are the shortest in length while technical reports are the longest. Generally, journal articles receive the most rigorous editorial review. Publications also differ in style. A technical report might have an executive summary that is not often seen in a journal article. Style elements depend on the publication. However, all scientific publications usually have four similar major components including an introduction, method, results, and discussion section.

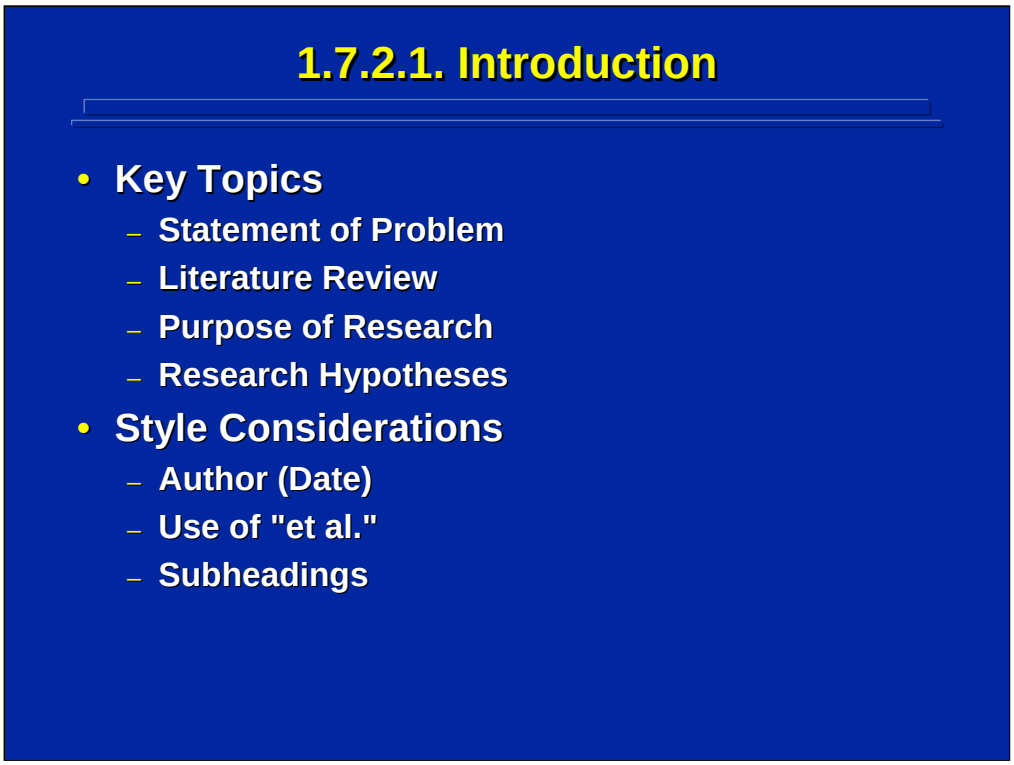



Scientific publication is characterized by objective reporting, and researchers' opinions are restricted to designated sections. The results section is the objective reporting of results and data analysis. The discussion section presents the researcher's opinions and interpretation of the results. There is often a restriction on length of the publication. Active voice is used to make it more interesting instead of a passive voice. Historically, the third person is used as opposed to first person. However, some journals are now allowing the use of first person narrative.

The general flow of the scientific report follows a story metaphor. Each section of the report helps tell the scientific story. There are four major components. Each section has a unique purpose, but these sections are integrated. The introduction tells readers the purpose of the research, its context in the scientific literature, and why they should read the report. The method tells readers how the data were collected and what constraints were set up to conduct the research. The results summarize the objective data and statistical analyses. In the discussion, the author explains and interprets the results. Additionally, the discussion ends the scientific story be returning to the purpose as stated in the introduction. Before writing a scientific story, one should outline the story line to insure integration of the report components.

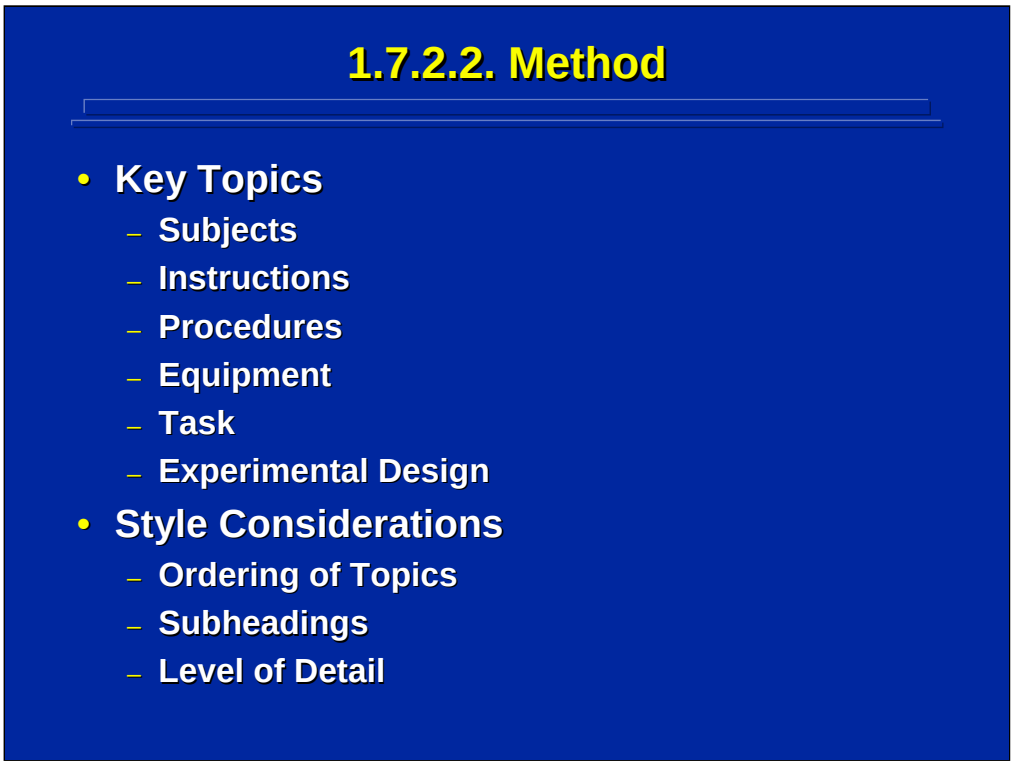



The next four slides summarize some of the major topics and key considerations of the introduction, method, results, and discussion sections, respectively, of scientific reports.



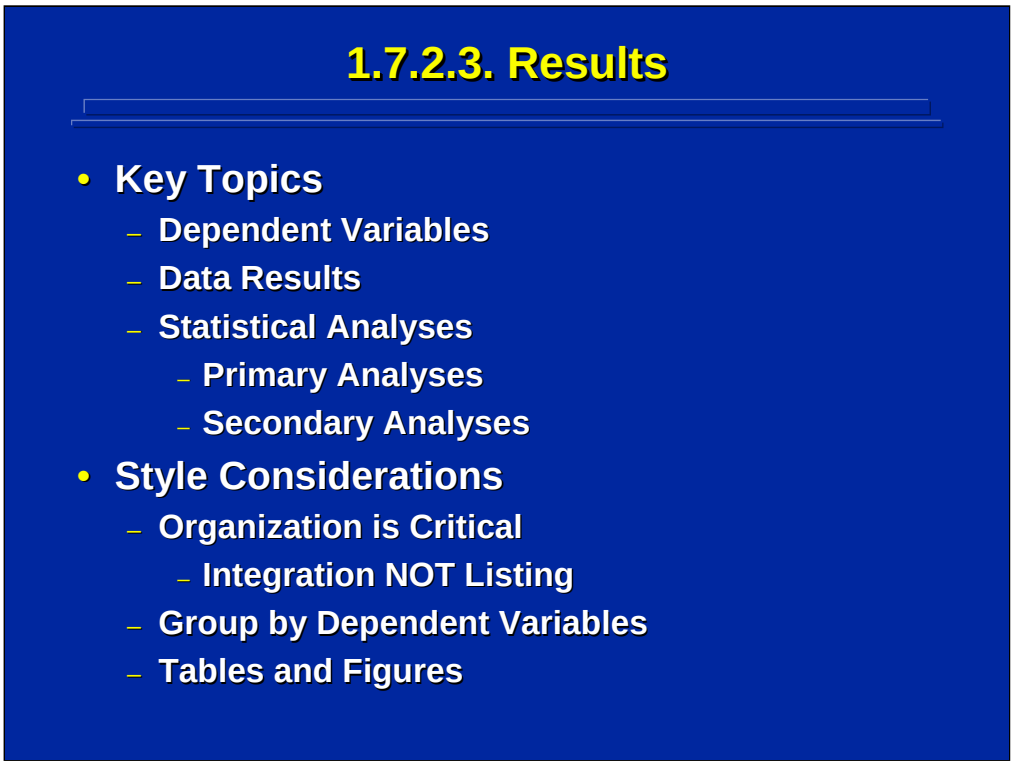

The introduction section should capture the reader's attention and state the problem in the context of the scientific literature. The purpose sometimes incorporates a hypothesis statement. Some type of literature review is also provided in this section. There are conventions for citing this literature. In the human factors literature, one usually makes literature citations by first stating the last name of the author followed by the publication date in parentheses. If there are several authors the "et al." statement can be used to eliminate repeating a list of authors after the first citation of all authors.

Subheadings can be used as a way of providing a road map for the reader. It is an easy way to guide the reader through the introduction from the literature review, to the purpose of the research, to the specific hypotheses being tested. Most journals allow several subheadings, and some journals allow the use of numbered subheadings to aid the reader.



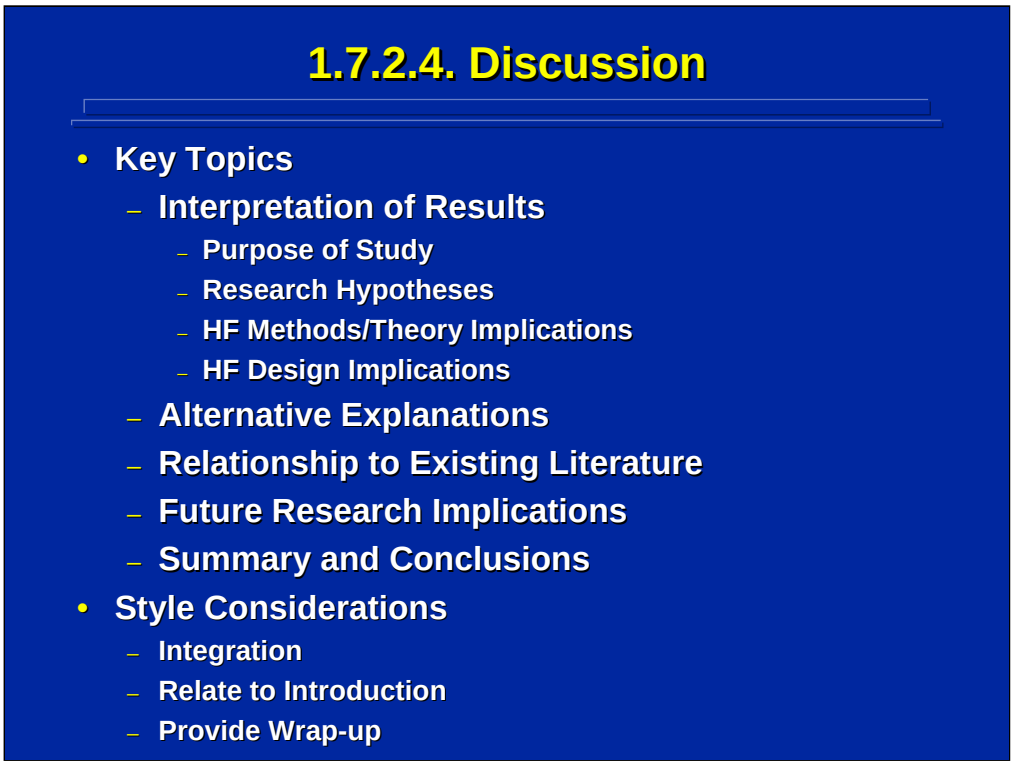

The method section should provide the reader with enough information to replicate the study. If there is not space for a detailed description, at least the critical aspects of the method should be presented. Some key topics on the method section are subjects, instructions, procedures, equipment, tasks, and experimental design.

The ordering of various topics in the method section is different for each report. Look for a good logical order for telling a scientific story. If one has to use words such as “to be discussed later” or “as stated earlier”, they are indications that the order is incorrect. The report should flow naturally. Subheadings can help guide the reader through the various components of the method section. The level of detail will depend on the study and space restrictions presented in publication guidelines.



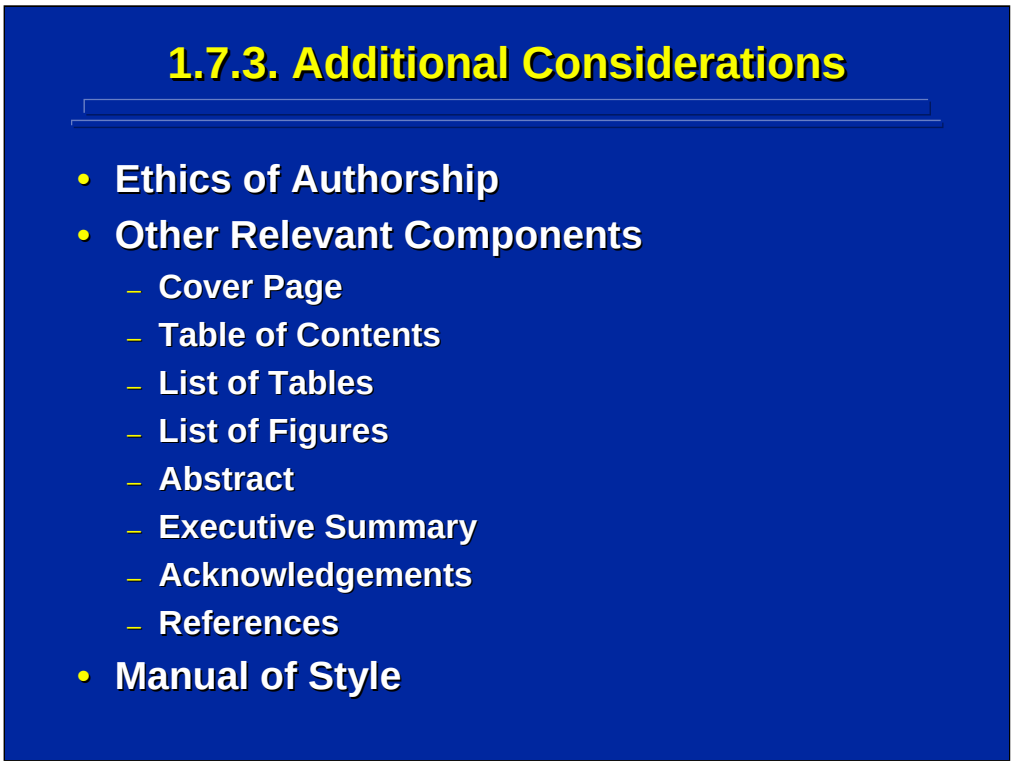

The results section should tell the reader what the dependent variables are, the measures taken while collecting the data, and the actual results. It should also include a summary of statistical analyses. The results are different from the statistical analyses. The results are stated descriptively in terms of means and variances of the treatment conditions; whereas, statistical analyses provide the tests of differences among the results. The actual differences stated in quantitative values (e.g., means) should be provided for all statistical tests to aid in the interpretation of the results. Statistical analyses can be divided into primary and secondary analyses. The primary analyses are those conducted on the dependent variables collected in the experiment, and the secondary analyses are conducted on follow-up data and questionnaires.

Organization of the results section is critical. Integrate the results into logical groupings. Just listing results can make this section boring and confusing to the reader. A common way to group results is by dependent variables. Another way to organize is around primary results and secondary results. Tables and figures should be used to make it easier for the reader to understand the results. The text should enhance, not repeat, the figure or table information. Remember, some publications restrict the number of tables and figures that can be used.




The discussion section should include experimenter interpretation of the results. Interpretations can be supported by the existing scientific literature, and references should be made where appropriate. Refer back to the purpose, problem, method, and results for implications. Also include alternative explanations of the results. These can lead to future research implications. Stating some conclusions at the end can be an effective way to wrap up the discussion section. In human factors research, a conclusion may also result in the statement of design guidelines based on the results of the experiment.

Usually one should keep the results and discussion sections separate. In a very complex experiment one might find a combined results and discussion section appropriate for improved communication to the reader. Remember that combining results and discussion also combines objective results with experimenter opinions. Style is based on how best to integrate the discussion. Subheadings can be used to aid in this integration. The scientific story is always a closed-loop story that should refer back to the purpose as stated in the introduction. Consequently, one should provide a wrap-up paragraph or sentence to avoid an abrupt ending.




The ethics of authorship are difficult. Usually authorship depends upon contributions to the actual scientific report writing, and the order of authorship reflects the level of written contribution. However, this does not always hold. There is no simple answer for authorship, and every researcher should develop personal guidelines for this decision.

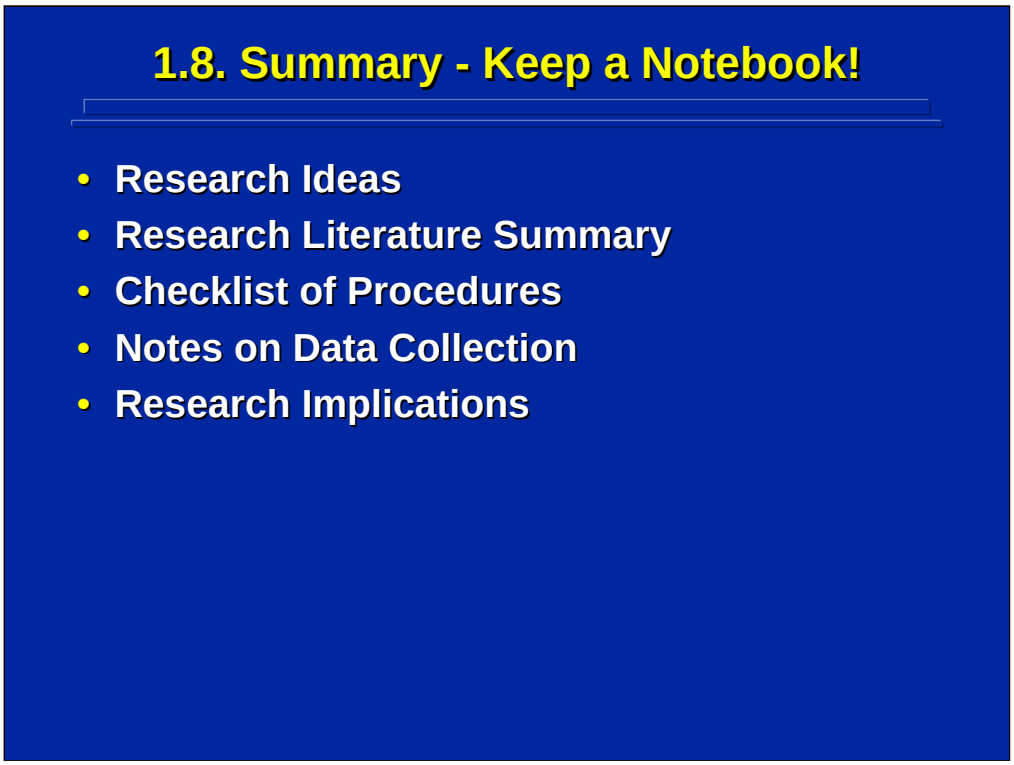
Other components of scientific reports may include a cover page, table of contents, list of tables, list of figures, abstract, executive summary, acknowledgements, and references depending on publication guidelines. For example, an abstract is quite useful in drawing attention to the report and in referencing it. Executive summaries are a four or five page summary of a long and detailed report. One should always try to include an acknowledgement section. This section recognizes others such as sponsors, software programmers, etc. who helped in the research but did not actually write the report. Finally, the reference list is important so that readers can refer to the other related articles.

The appropriate manual of style should always be considered in preparing a scientific publication. Human factors researchers generally use the American Psychological Association (APA) Manual of Style that is discussed by Martin (2004).

## 1.8. Summary - Martin's Critical Questions

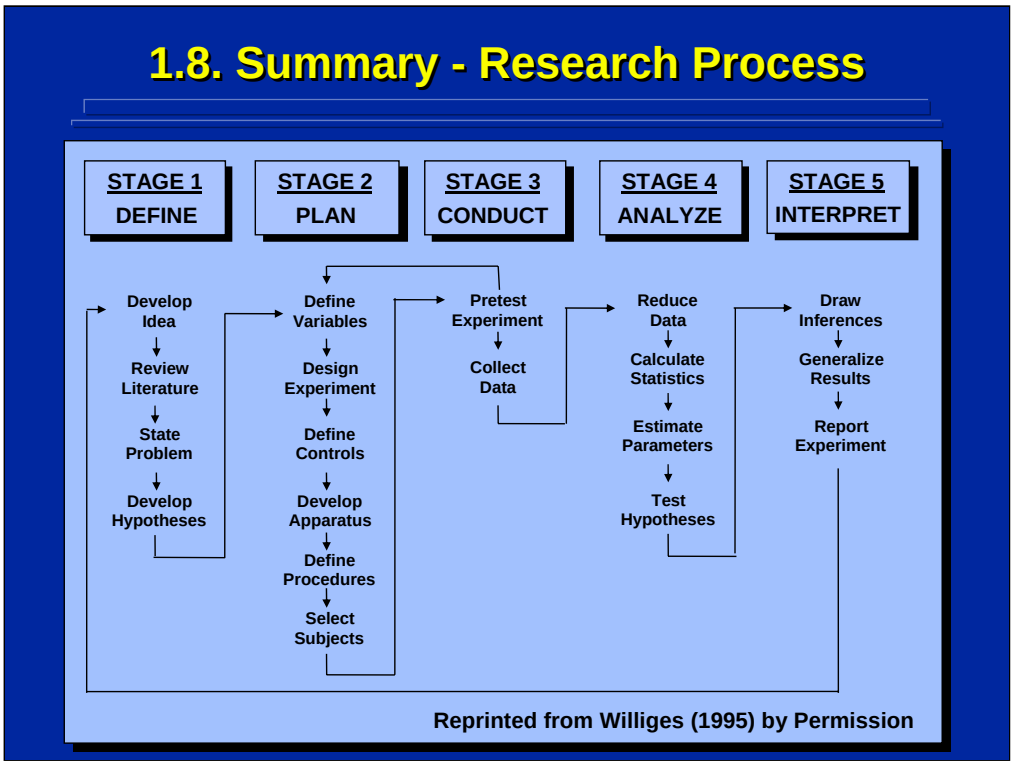
- Does my experiment satisfy ethical concerns?
- How many subjects do I need?
- Should I run subjects individually or in groups?
- How long will my experiment take?
- Do I need to set subject restrictions?
- Should I set any a priori criteria for eliminating subjects?
- Can I operationally define all my variables?
- Have I arranged for any equipment or materials needed?
- Do I know how I will analyze my data?
- How will I interpret the possible outcomes of my experiment?

Martin (2004) on pp. 233-242 discusses each of these critical questions to ask oneself before data collection. Experimenters should develop a similar checklist before collecting data that is tailored to their research problem area.

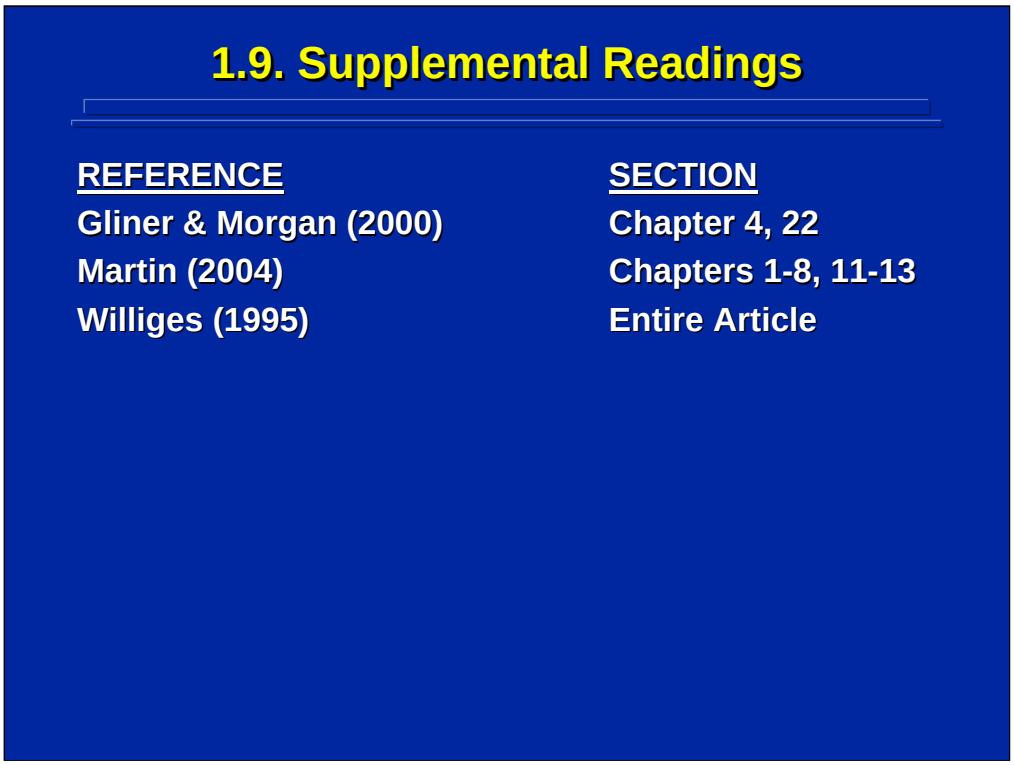



Keep a research notebook! This is one of the most important aspects of research, like, pretesting, that is often overlooked. Several items might be included in a researcher's notebook. It might include research ideas that could lead to future related experiments. A research literature summary should be kept including the complete reference citation. Notes may also be used to develop a checklist of procedures before data collection.

The experimenter should make notes during data collection to document possible outliers and unusual circumstances such as equipment failures that could affect the results. While collecting data, the researcher should keep notes on possible outcomes and implications for interpreting the results. One should be compulsive in keeping notes on items that may be difficult to remember. Taking notes throughout the research process can facilitate writing the subsequent research report.

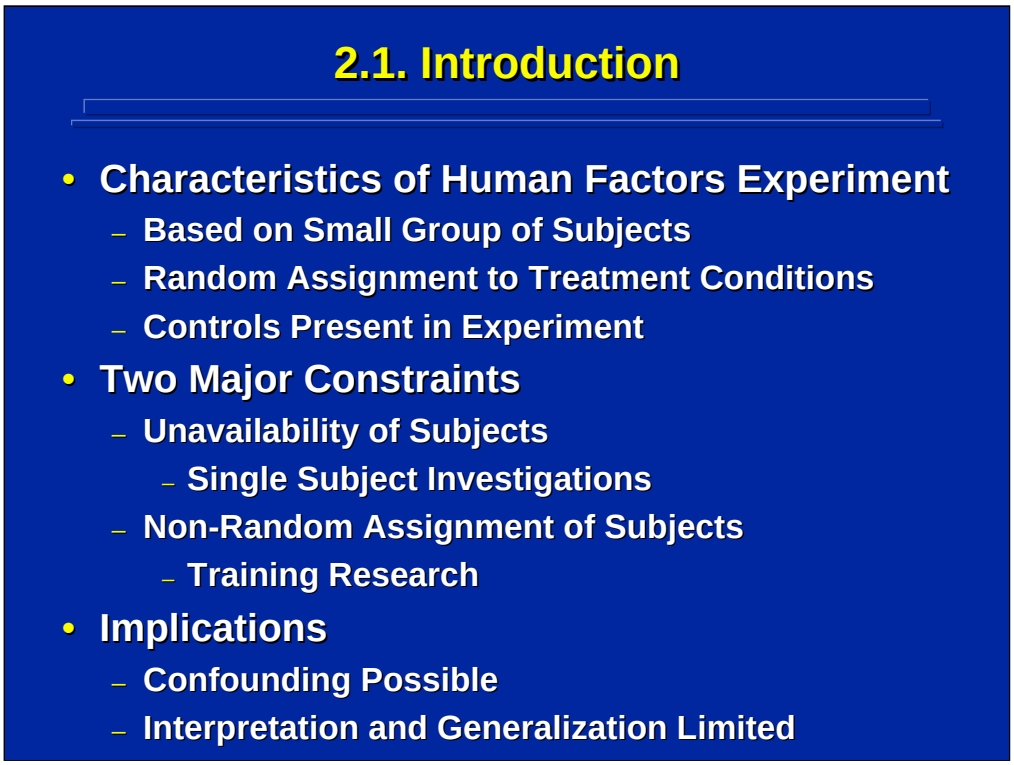



By way of summarizing the research process, refer once again to the five stage research process diagram developed by Williges (1995). (This figure is reprinted by permission of Person Education, Inc., Upper Saddle River, New Jersey.) Notice that it is a closed-loop process involving many considerations besides the critical items covered in this topic. New research implications from the results of one experiment can lead to restarting the research process on a related problem.



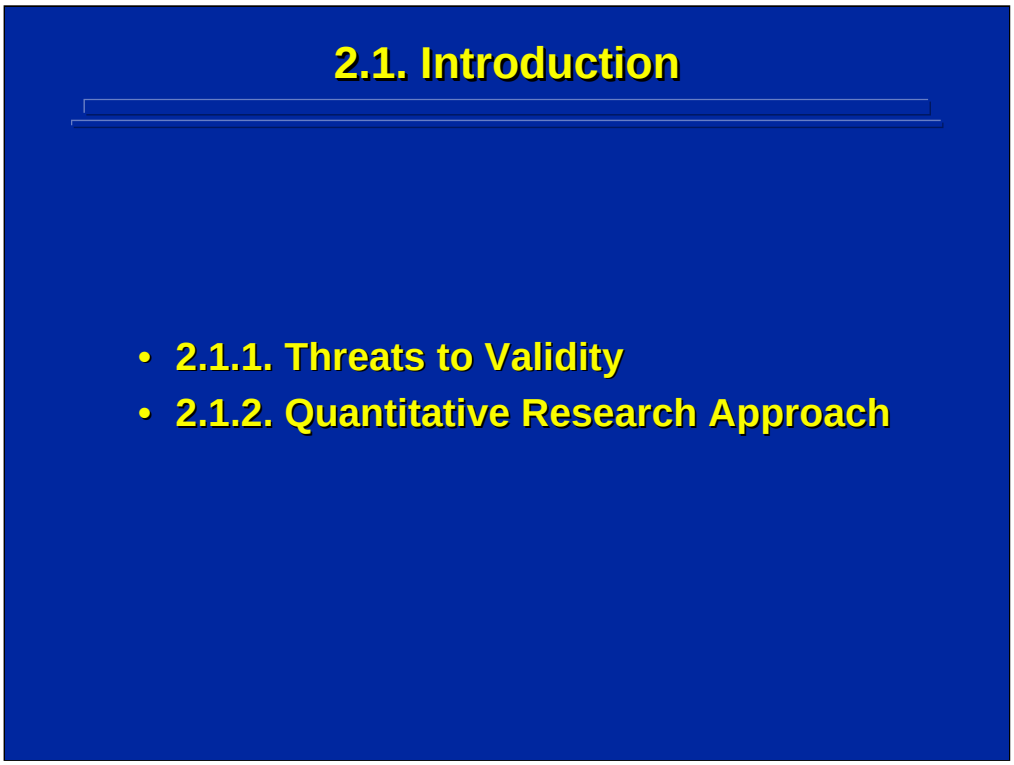

The Martin (2004) reference is specifically directed toward conducting research on human subjects. It is a highly entertaining treatment of the basic components involved in conducting research and is highly recommended for human factors researchers who have had little or no experience in collecting data from human subjects. The Williges (1995) book chapter expands on the various concepts depicted in the research process flow diagram of the five stages of research as presented in this reference material. Gliner and Morgan (2000) discuss, the choice of research questions and variables, the appropriate treatment of human subjects, and ethical issues related to authorship.

This topic is an introduction to experimental design. It begins with an overview of various conditions that can threaten the validity of the results of experiments and then offers a general notation for designating designs. Two general experimental design alternatives are presented, and the case is developed for always choosing randomized experimental designs, if possible.

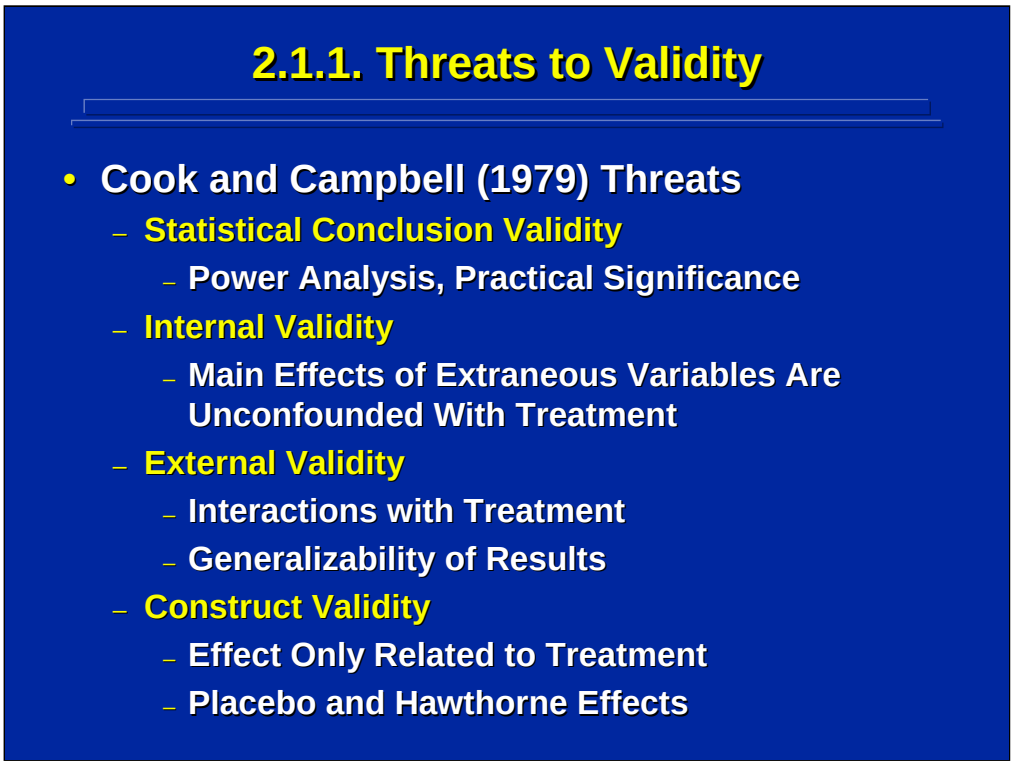



Human factors experiments generally are characterized by having a small group of subjects, random assignment of subjects to treatment conditions tested in the experiment, and control of other factors (i.e., instructions, testing times, etc.) that might influence the data.

Two common constraints that often occur in human factors experiments are unavailability of a large sample of subjects and the inability to randomly assign subjects to treatment conditions due to real world settings. For example, subjects for a training research experiment cannot be randomly assigned to treatment conditions since they are already assigned to classes that cannot be divided. These constraints lead to confounding which can limit interpretation and generalization. This reference material provides some design alternatives to deal with these two constraints.



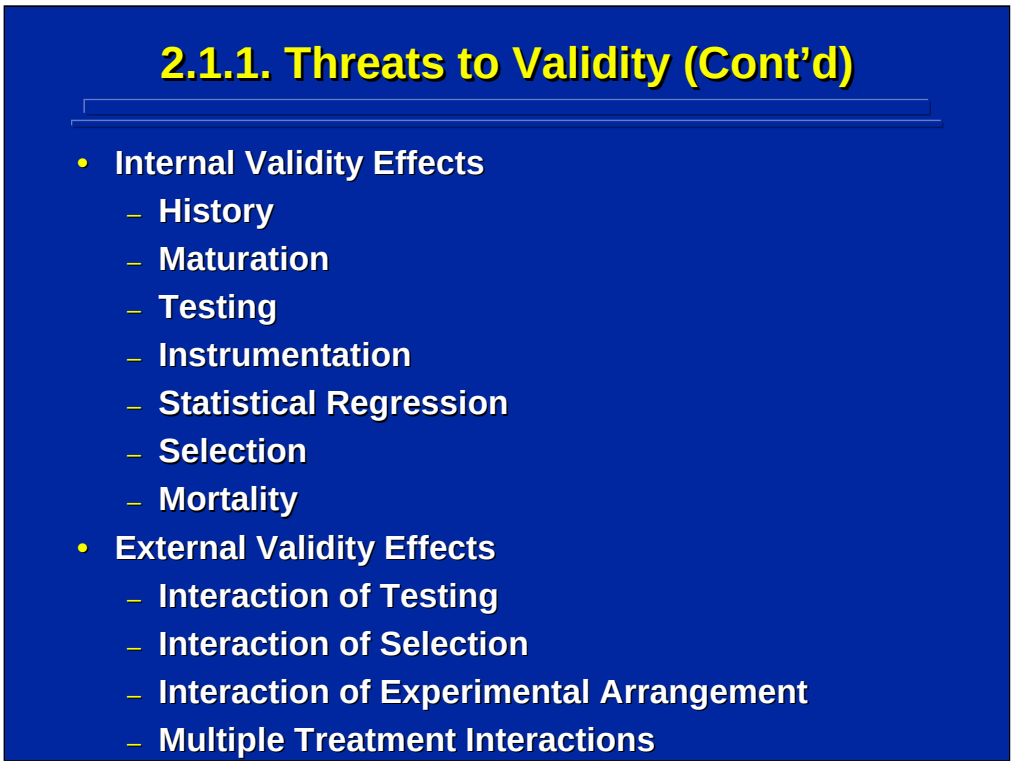

Various factors can limit interpretation of data from experimental designs, and steps can be taken to help control these threats to validity. These controls require manipulation of conditions through experimental designs yielding quantitative results rather than passive observations.




Cook and Campbell (1979) describe four specific threats to the validity of interpretation of data collected from experiments in Chapter 2. Shadish, Cook, and Campbell (2002) provide an extended discussion of how these threats to validity affect causal inferences based on research results in Chapters 2 and 3.

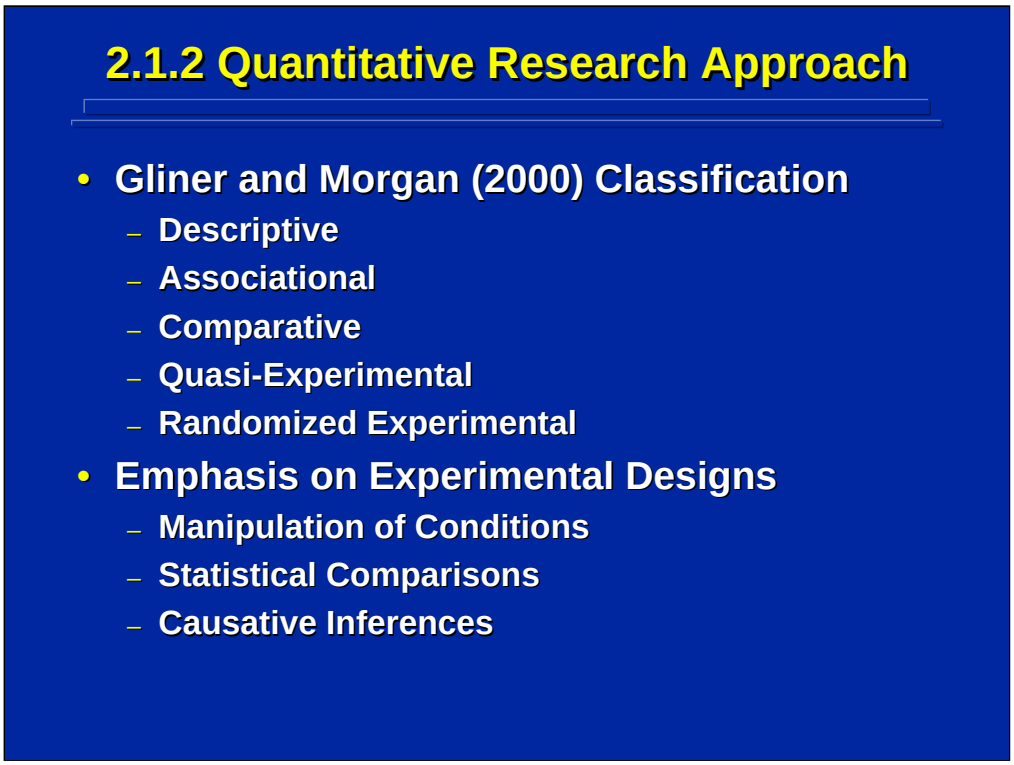
Statistical conclusion validity deals with the power of the test and the practical significance of the test. Statistical power is the ability to find a true difference if a true difference exists. Experimenters strive to use the most powerful tests possible for their research. Generally, a randomized experimental design provides data for the most powerful test. Just because a result is statistically significant, it may have no practical significance in terms of interpreting the results in real-world applications. Human factors researchers are usually interested in statistical differences that also have practical significance.

Internal validity deals with keeping main effects of the experiment separate from other confounding factors that may affect interpretation. External validity is the interaction of these confounding factors with the factors manipulated in the experiment. That interaction may affect the generalizability of the results. Construct validity is compromised when another factor such as previous testing is really causing the difference, not the construct being tested. Of these four types of threats to validity the two most important in designing an experiment are internal and external threats.



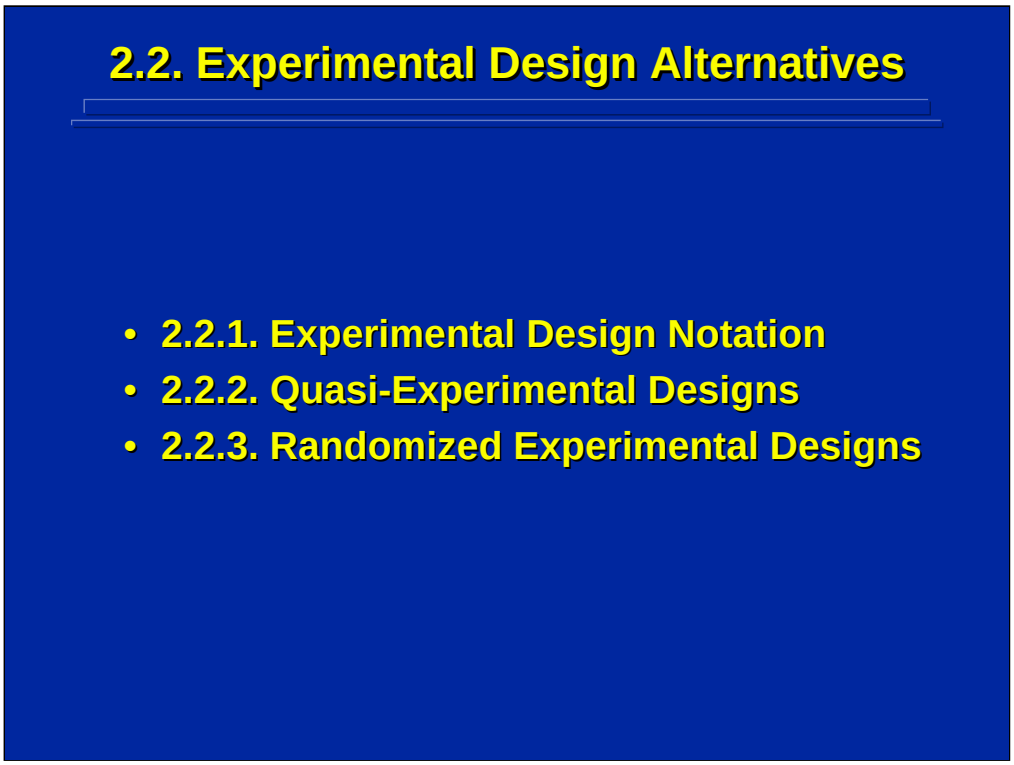

Cook and Campbell (1979) describe six internal threats that can confound results. For example, mortality occurs if a subject drops out of the experiment before completion. Certain groups of people could become bored or frustrated with the test and drop out. Other internal threats are history, maturation, testing, instrumentation, and statistical regression. See pages 50-58 of Cook and Campbell (1979) for a detailed discussion of these internal threats.

External validity threats are the interactions of the internal validity threats with the treatment conditions of interest. Interactions of testing, selecting, and experimental arrangement are common examples of external threats. Experimenters should consider the possible internal and external threats when choosing an experimental design in order to improve interpretations and generalizability.

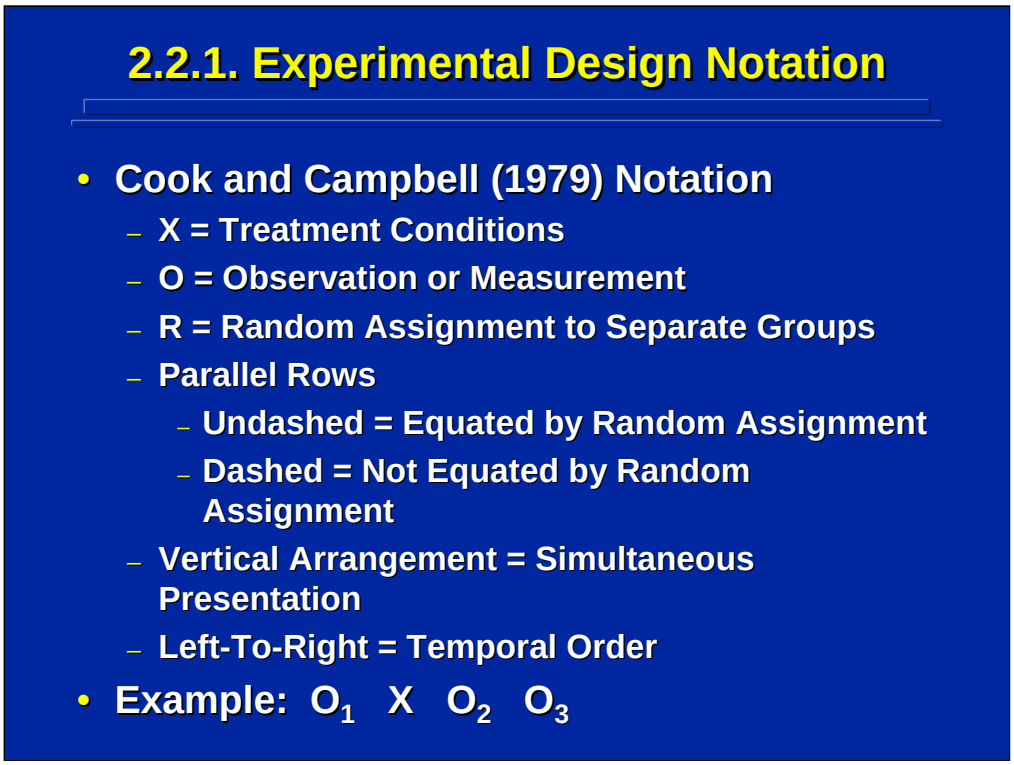



Gliner and Morgan (2000) in Chapter 5 classify research into five different quantitative approaches that explore relationships among variables. These approaches vary in purpose and criteria required in using each approach. Descriptive and associational approaches are not concerned with causative inferences. The comparative approach is concerned primarily with differences between groups. Both the quasi-experimental and the randomized experimental approaches are concerned with causality.

Since human factors researchers are primary concerned with comparing conditions to make causative inferences, they usually employ some type of experimental design in conducting their research. These designs are used to manipulate variables of interest, collect quantitative data for statistical analysis, and control confounding variables as much as possible that might confound causative interpretations.




Cook and Campbell (1979) wrote the definitive treatment on quasi-experimental designs. Their design notation is used to compare two general classes of experimental designs: quasi-experimental designs and randomized experimental designs. Various considerations are presented for both types of experimental design in order to demonstrate the strength of randomized designs that are preferred in human factors research.



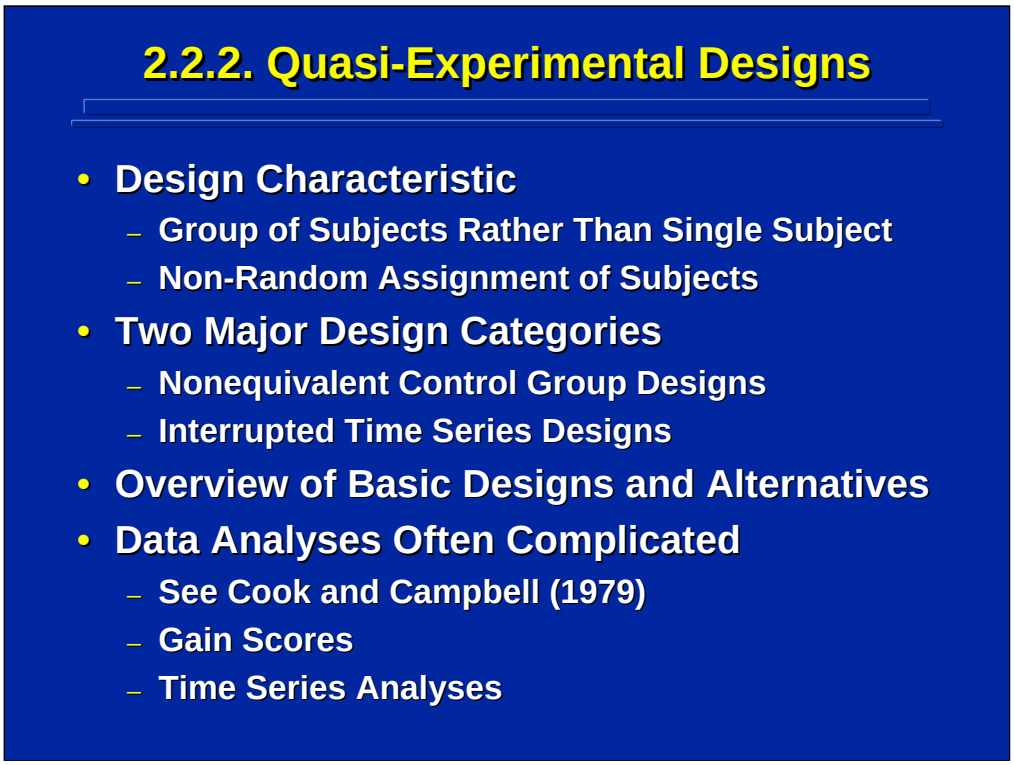

Cook and Campbell (1979) use the following notation for laying out experimental designs. An “X” is a treatment condition, or an independent variable. An “O” is an observation, measurement, or a dependant variable. The “R” designates random assignment of subjects to separate groups.

Various treatment combinations are presented in parallel rows. If there are no dashes between the rows, then subjects are randomly assigned. If there are dashes between the rows, then the subjects are not equated by random assignments. Everything that appears in the same column is presented simultaneously. Temporal order goes from left to right.

This notation is used to describe the designs listed throughout the remainder of this topic. For example, designation of a treatment condition in a training experiment might be:

$O_1$ X $O_2$ $O_3$

where $O_1$ is a pretest measure before training followed by administration of the experimental training condition, X, then performance on the first practice trial, $O_2$, and finally performance on the second practice trial, $O_3$.

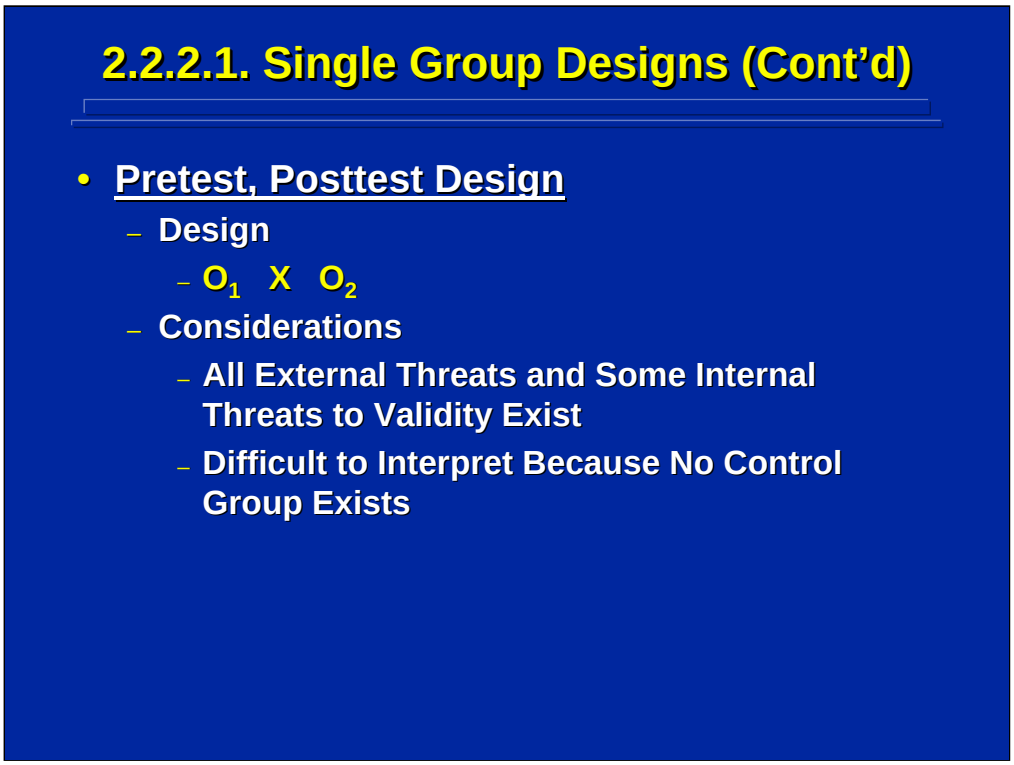



Quasi-experimental designs usually use a sample of subjects rather than a single subject, include a treatment condition, and may include control conditions. The key characteristic of a quasi-experimental design, however, is the non-random assignment of subjects to treatment conditions. If one could randomly assign subjects to treatment conditions, then one would have a randomized experimental design.

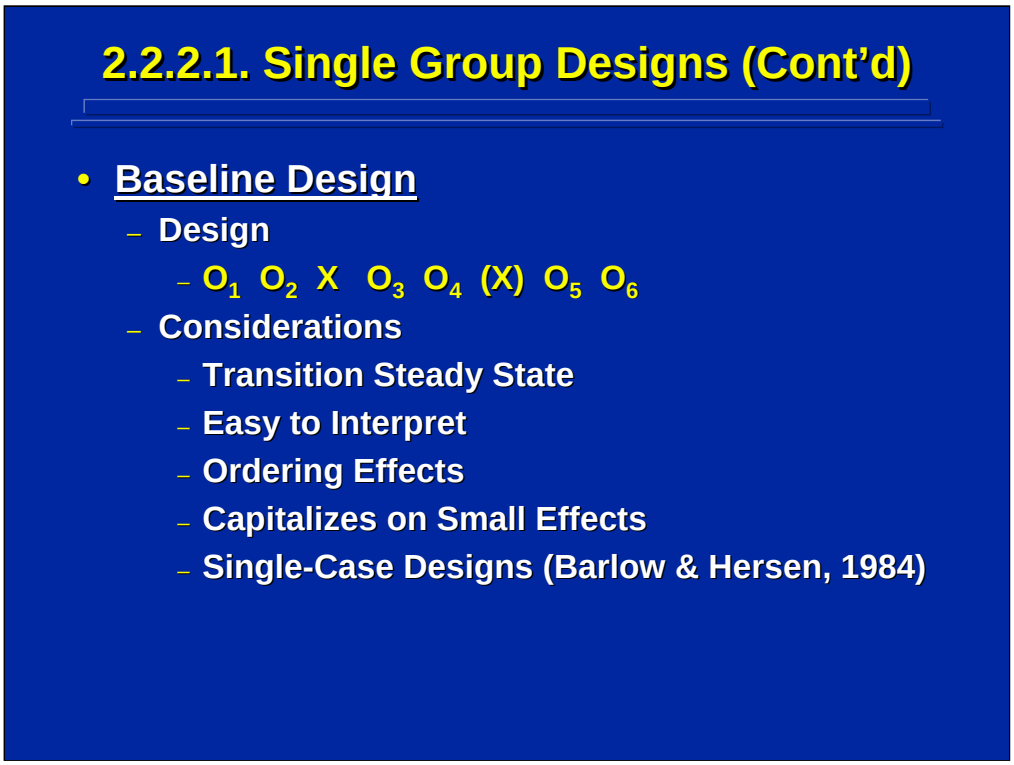
The two major categories of quasi-experimental designs are nonequivalent control group and interrupted time series designs. If one uses a non-equivalent control group design, one would often use gain scores in the analysis. If one chooses an interrupted time series design, one would use a time series analysis. Only a few examples of quasi-experimental designs presented by Cook and Campbell (1979) and Shadish, Cook, and Campbell (2002) that are useful in human factors research are presented in this reference. No discussion of detailed data analysis techniques for any of these design examples will be provided since the emphasis of this reference material is on true experimental designs.

Shadish, Cook, and Campbell (2002) in their Chapter 4 consider experimental designs that lack pretest observations or do not have a control condition as a separate group of quasi-experimental designs. Gliner and Morgan (2000) consider these single group designs as weak quasi-experimental designs because many internal and external threats to validity exist.

A classic single group design is commonly known as the case study or the one shot design. All of the internal and external threats to validity exist. This design is usually characterized by naturalistic or direct observation, O, in a very specific situation, X. A case study is used when experimental variables cannot be manipulated in the real world. In human factors research, a case study involving field operation can provide preliminary data for designing a subsequent randomized experiment.




Another type of single group quasi-experimental design is the pretest, posttest design. First, baseline data, $O_1$, are collected before a treatment condition, X. Then, more data, $O_2$, are collected after the treatment is given. For example, such a design might be used in field studies dealing with motion sickness. The before and after treatment data are compared to document the treatment effect. Since there is no control condition, it is difficult to determine whether the difference is due to the treatment or just practice.

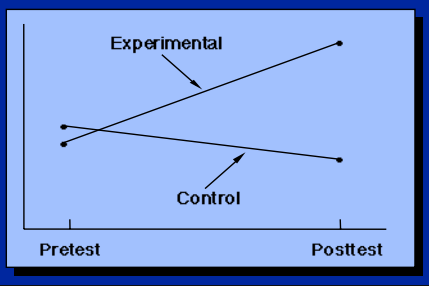



A baseline design is another type of single group design. In this research design several observations, $O_1$ and $O_2$, are made before the treatment, X, followed by several observations, $O_3$ and $O_4$, after the treatment. Then there may also be a period where nothing happens shown by "(X)" followed by more observations, $O_5$ and $O_6$.

A baseline study allows one to analyze transitions from steady state performance. One hopes to see a break, or a jump, in performance after the treatment occurs. This design is often used to control ordering effects. Since there is no control condition, one must infer the control condition by way of multiple observations on a single group. The Barlow and Hersen (1984) text on single-case designs shows various alternative baseline designs that may be useful to human factors researchers who need to conduct single-case studies.

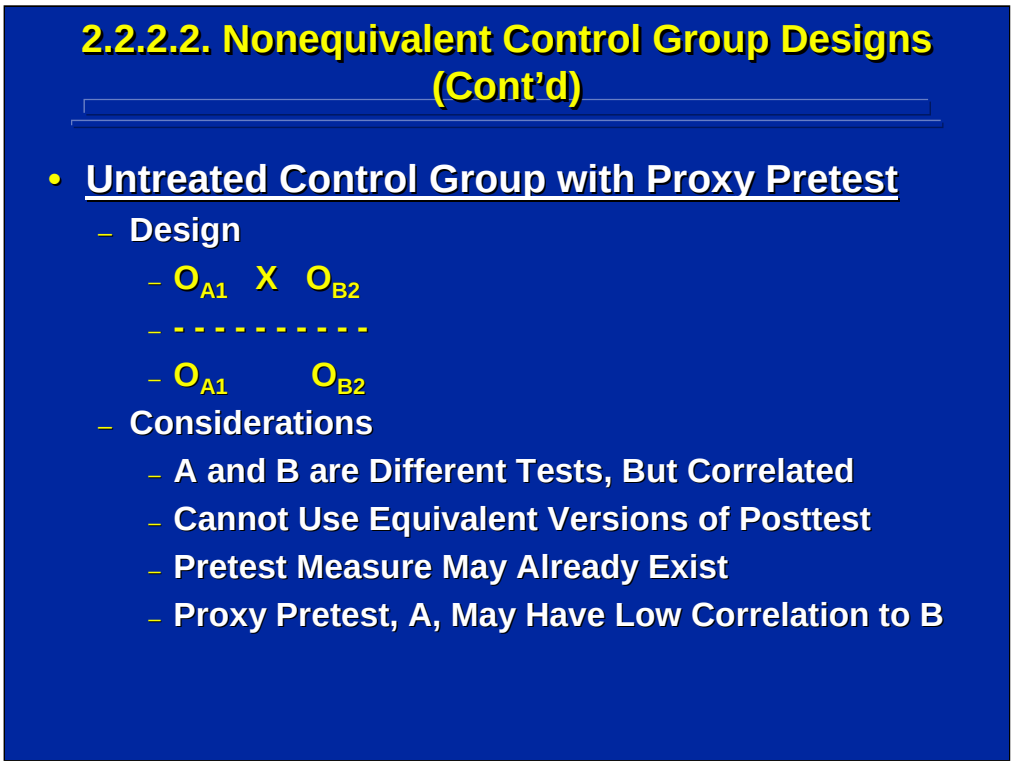
## 2.2.2.2. Nonequivalent Control Group Designs

- **Basic Design: Untreated Control Group with Pretest and Posttest**
  - **Design**
    - $O_1$ X $O_2$
    - - - - - - - - -
    - $O_1$ $O_2$
  - **Ideal Outcome**




The basic nonequivalent control group design is called untreated control group with pretest and posttest, because the control group only has a pretest observation, $O_1$ and a posttest observation, $O_2$, with no treatment, X, as shown on the slide. The key difference between this design and a randomized experimental design is non-random assignment of subjects as indicated by the dashed line shown on this slide.

The ideal outcome according to Cook and Campbell (1979) is performance improvement from pretest to posttest only in the experimental condition and not in the control condition as shown on the slide. If this result does not occur, interpretation is difficult since the non-significant effect could represent internal threats resulting from non-random assignment of subjects. In general, quasi-experimental designs yield straightforward interpretations if the expected outcome occurs, but non-expected outcomes may be difficult to interpret.



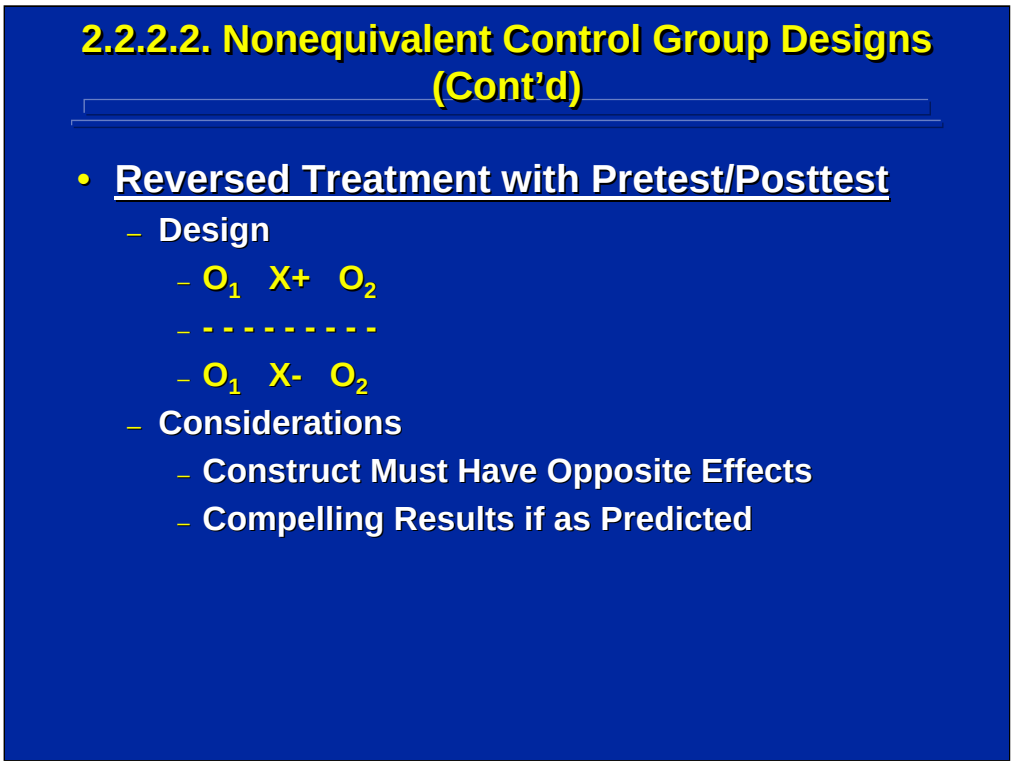

A variation to the basic nonequivalent control group design is the untreated control group with a proxy pretest, $O_{A1}$. This alternative is useful when equivalent tests cannot be generated for pre-testing and post-testing. As shown on the slide, each group gets both tests A ($0_{A1}$) and B ($0_{B2}$). Although tests A and B are different, they are correlated. If the two tests were not significantly correlated, then test A might not be an appropriate pretest for Test B.

## 2.2.2.2. Nonequivalent Control Group Designs (Cont'd)

- **Nonequivalent Dependent Variable Design**
  - Design
    - $O_{1A}$ X $O_{2A}$
    - - - - - - - - - - - -
    - $O_{1B}$ (X) $O_{2B}$
  - Considerations
    - Different Measures on a Single Group of Subjects
    - Assumes B is Not Affected by Treatment, X
    - Can Be Extended to Additional Tests, C ... N
    - Tests Must be Conceptually Related
    - Can Use with Other Designs

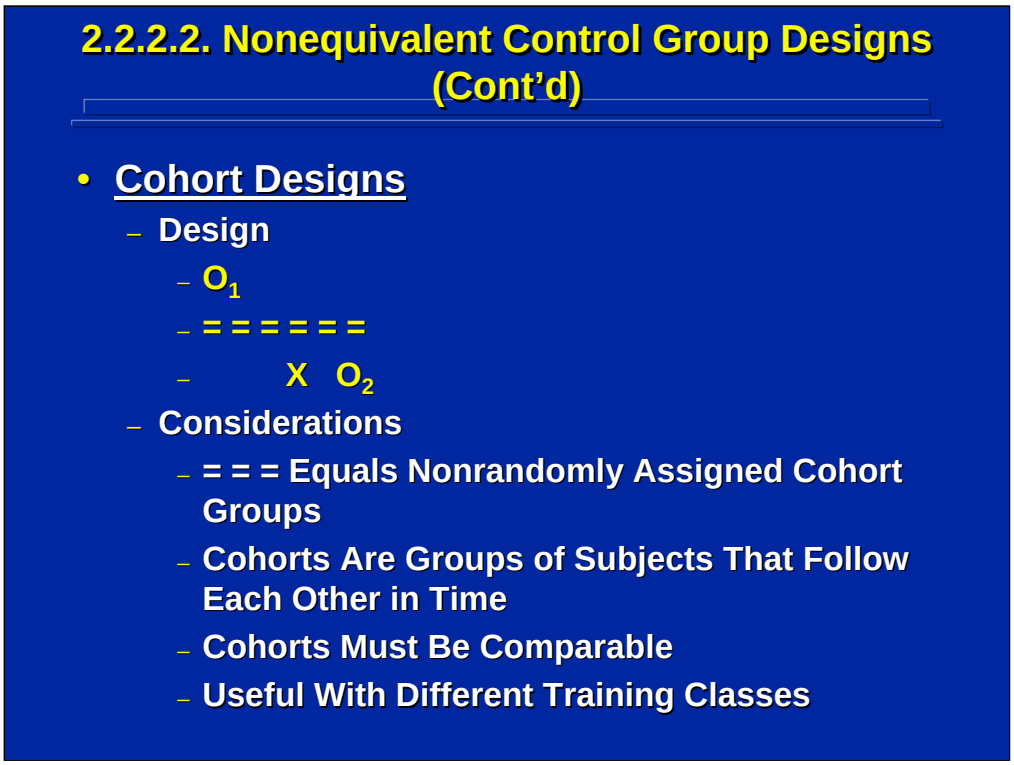
Another alternative to the basic quasi-experimental design is to use a nonequivalent dependent variable design. Two different conceptually related tests A and B, such as verbal comprehension and reading speed, are obtained from only one group of subjects in both the pretest and the posttest. Only test A should be affected by the treatment condition X; whereas test B should not be as denoted by "(X)" on the slide.

This is a useful alternative when subject availability is limited and not randomly assigned as denoted by the dashed line on the slide. Nonequivalent dependent variables also can be used in conjunction with other quasi-experimental designs to control for retesting effects.




A reversed treatment with pretest/posttest is a third alternative to the basic quasi-experimental design. With this alternative, two different treatments are tested that have predicted opposite performance effects (i.e., X+ and X-).

If $0_2$ performance improves as compared $0_1$ in one condition (X+) and deteriorates in the other condition (X-) as predicted, then the results are compelling. However, if the results are not as predicted, the outcome can be due either to no treatment effect or to validity threats resulting from non-random assignment of subjects to the two conditions as indicted by the dashed line on the slide.




Cohort designs use non-randomly assigned cohort groups. Cohort groups are nonrandom groups of subjects that have many similarities and are comparable, but follow each other in time. For example, different training classes of basic military recruits are cohorts, because selection criteria result in many similarities among recruits.

Cohort designs are particularly useful in training experiments where subjects cannot be randomly selected for different treatment conditions but must be assigned by classes that follow each other in time. As shown on the slide the first cohort group is observed, $O_1$, and provides the pretest score for a second cohort group that follows the first group and receives a treatment, X, followed by a posttest, $O_2$.

## 2.2.2.3. Regression-Discontinuity Designs

- **Regression-Discontinuity Design**
  - Design
    - $O_1(C)$ X $O_2$
    - ----------
    - $O_1(C)$ $O_2$
  - Considerations
    - $O_1$ is a Continuous Pretest Measure
    - X Presentation Depends on Cutoff Value (C) of $O_1$
    - $O_1$ and $O_2$ Must be Highly Correlated
    - Assumes Linear Slopes

Shadish, Cook, and Campbell (2002) devote an entire chapter, Chapter 7, to regression discontinuity designs as quasi-experimental designs that are completely separate from nonequivalent control group designs due to the unique establishment of the control group by the researcher. The control group is created based on a cutoff value (C) of the continuous pretest measure, $O_1$, set by the experimenter. The treatment, X, is only presented to subjects who score above the C value of $O_1$. If subjects score at or below the C value on $O_1$, they become members of the control group and do not receive the treatment.

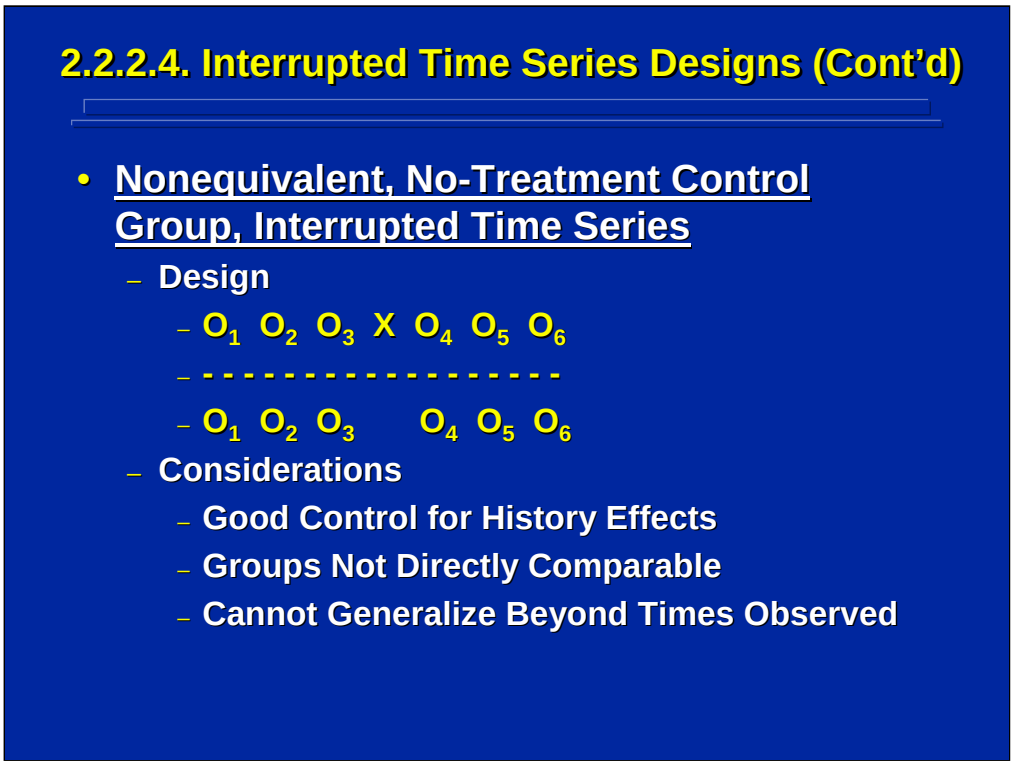
The $O_1$ and $O_2$ observations are highly correlated resulting in a linear increase when they are plotted together for both the treatment and non-treatment group of subjects. The effect of the treatment, however, should result in a jump, or discontinuity, in the linear increase between $O_1$ and $O_2$ only for the treatment group and not the control group. The amount of discontinuity is interpreted as the treatment effect.

## 2.2.2.4. Interrupted Time Series Designs

- **Basic Design: Simple Interrupted Time Series**
  - Design
    - $O_1$ $O_2$ $O_3$ $O_4$ X $O_5$ $O_6$ $O_7$ $O_8$
  - Considerations
    - Treatment, X, Interrupts Series of Observations, O's
    - Need Only One Group
    - Could Use Archival Data
    - Directly Measures Maturation Effects

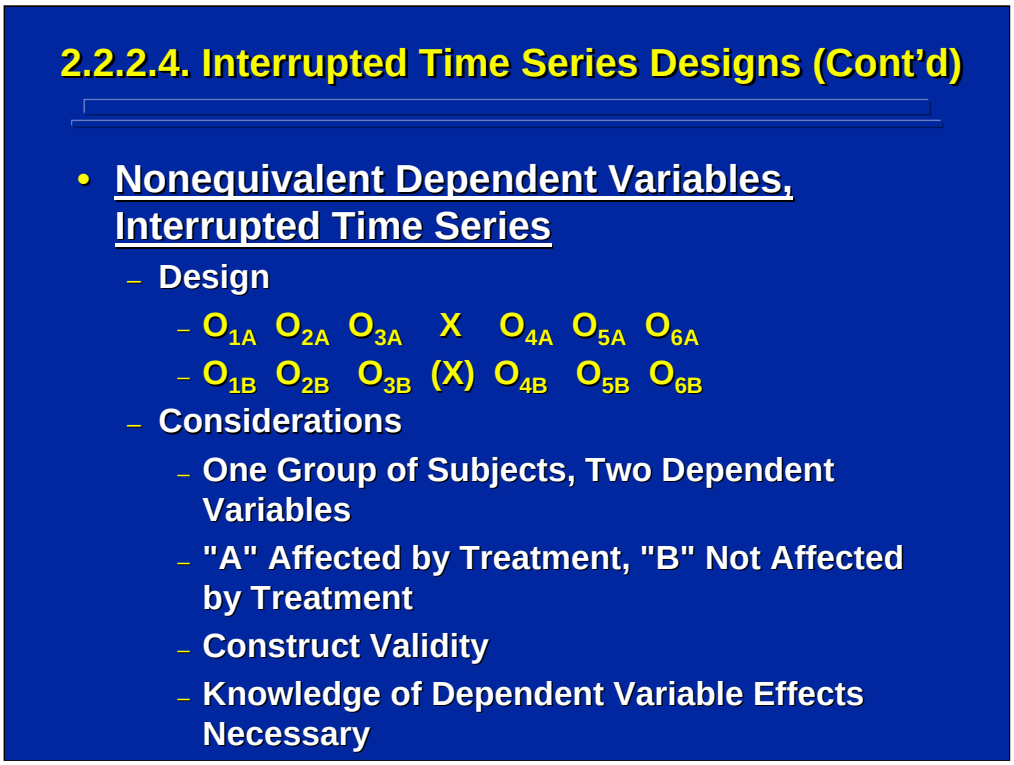
An interrupted time series design is another major class of quasi-experimental designs that are used when many observations can be taken over a period of time on a group of subjects. These observations may exist as archival data that are collected on a regular basis. The more observations that can be collected before and after the treatment, the more stable and robust the effects. As shown on this slide, the treatment, X, is presented to one group of subjects at a certain point in the pretest observation sequence (i.e., $O_1$ to $O_4$) to "interrupt" the time series of posttest observations (i.e., $O_5$ to $O_8$).

In the human factors literature there are very few examples of interrupted time series designs. Consequently, only two alternatives to the basic interrupted times series design are presented to show various possible quasi-experimental design alternatives. See Cook and Campbell (1979), Chapter 5, and Shadish, Cook, and Campbell (2002), Chapter 6, for a complete discussion of the various alternatives.




The basic interrupted time series used in quasi-experiments is the non-equivalent, no treatment control group design. Each group consists of non-randomly assigned subjects as denoted by the dashed horizontal line. The first group receives the treatment, X, during the observation series, $O_1$ to $O_6$; and the second group serves as the control that does not receive the treatment during the observation series.

This design is the time series equivalent to the basic quasi-experimental design. Comparison between the two groups is made in terms of measures related to the slopes and intercepts of regression analyses. Box and Jenkins (1976) in Chapter 7 and Cook and Campbell (1979) in Chapter 6 provide details on conducting time series analyses.



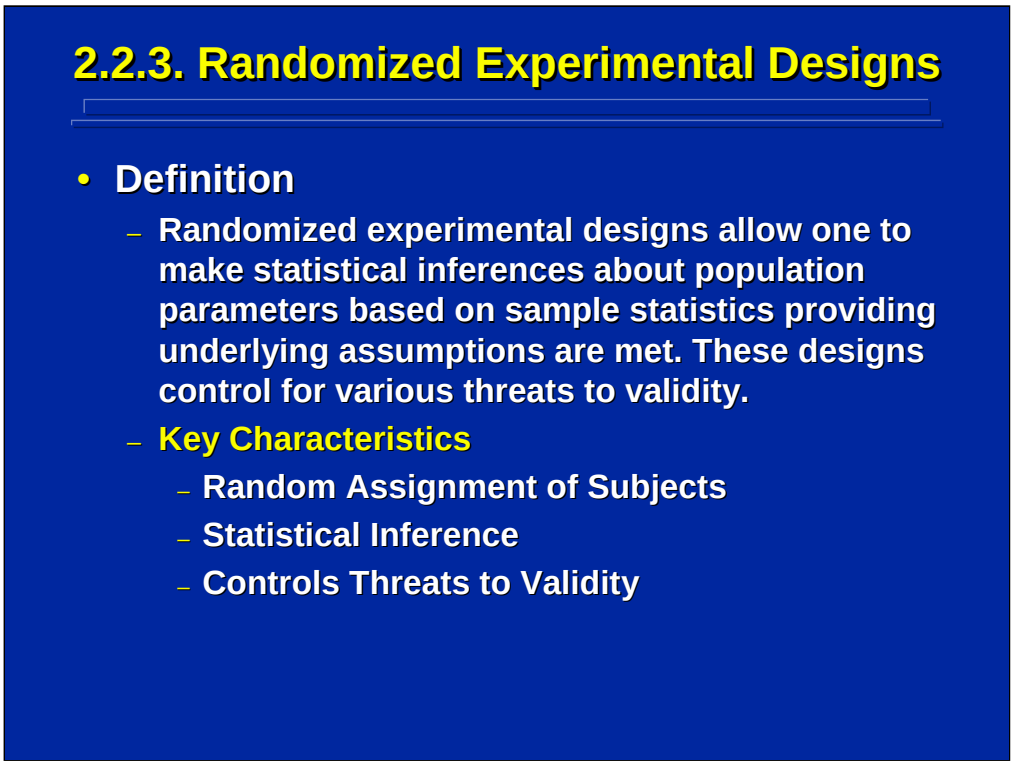

The time series equivalent to the non-equivalent dependent variable quasi-experimental design alternative is shown in this slide. Note that this design uses only one group of subjects who are measured repeatedly on both A and B. The treatment affects only measure A. Measure B serves as the control measure, because it is not affected by the treatment as shown by "(X)" on the slide. The two metrics $O_A$ and $O_B$ must be conceptually related (i.e., measures of two verbal and motor skills) for construct validity even though only one is affected by the treatment.

## 2.2.2.4. Interrupted Time Series Designs (Cont'd)

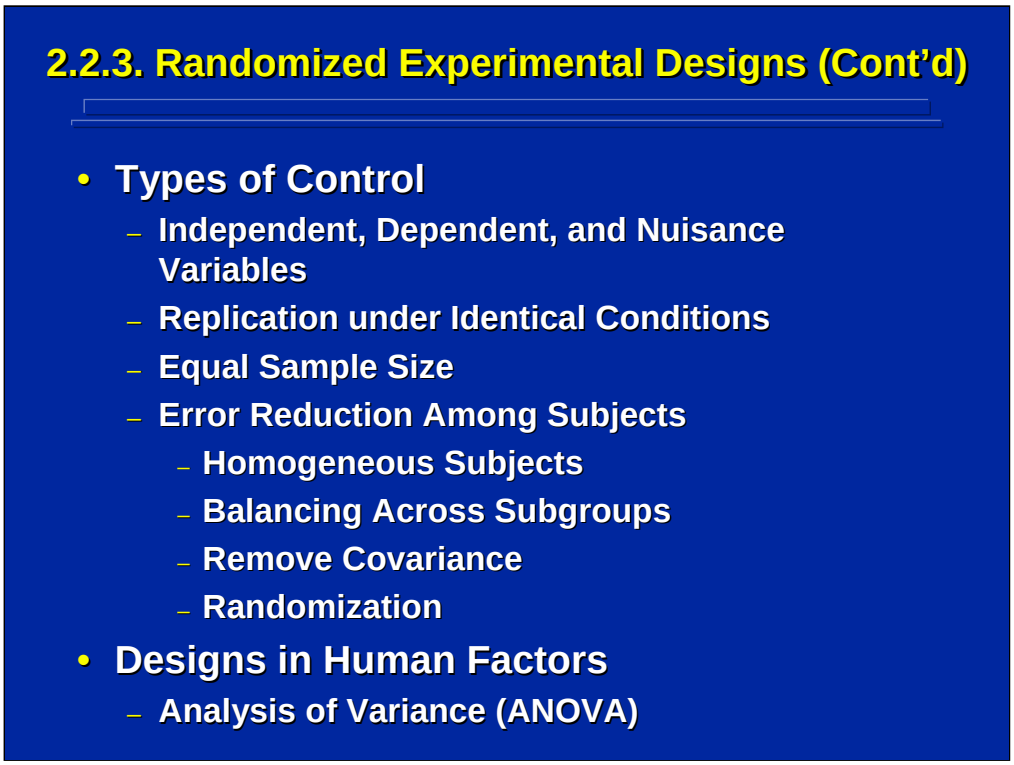
- **Switching Replications, Interrupted Time Series**
  - **Design**
    - $O_1\ O_2\ O_3\ X\ O_4\ O_5\ O_6 \quad O_7\ O_8\ O_9$
    - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
    - $O_1\ O_2\ O_3 \quad O_4\ O_5\ O_6\ X\ O_7\ O_8\ O_9$
  - **Considerations**
    - **Analyze as Two Separate Nonequivalent Control Groups**
      - $O_1 ... O_6$
      - $O_4 ... O_9$
    - **Each Group Serves as a Control for the Other**
    - **External Validity Enhanced With Two Comparisons**

The final interrupted time series alternative is the switching replications quasi-experimental design shown in this slide. Two groups of non-randomly assigned subjects are used. Each receives the treatment, X, but at different points in the time series. Consequently, two pairs of time series are compared as in the basic interrupted time series design.

The first pair of time series involves observations $O_1$ to $O_6$, and the second pair of time series involves observations $O_4$ to $O_6$ as shown on the slide. In one case, one group is the treatment condition but, in the other case, it becomes the control condition, thereby switching replications. This design alternative provides some external validity in terms of which group is the control and which is the experimental condition.



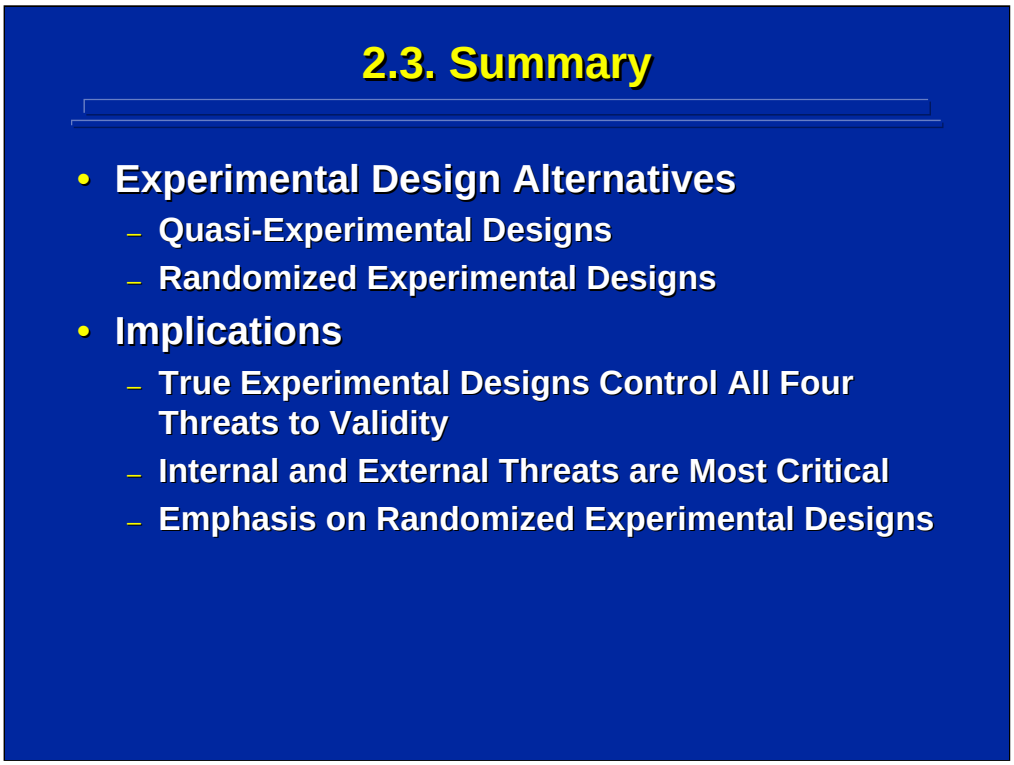

A randomized experimental design includes control conditions and random assignment of subjects. These designs provide controls for various threats to validity and allow statistical inferences about population parameters based on sample statistics providing that certain assumptions are met. Conversely, quasi-experimental designs are hampered by internal and external threats to validity, because they do not have random assignment of subjects. Consequently, randomized experimental designs always provide the best design alternative and should be used whenever possible in human factors research.




Randomized experimental designs have several major characteristics. They are specified in terms of independent and dependent variables and provide controls for nuisance variables such as different experimenters, testing times, etc. Randomized experimental designs have replications under identical conditions for every subject. Equal sample sizes are used for robustness to violations of assumptions of statistical tests.

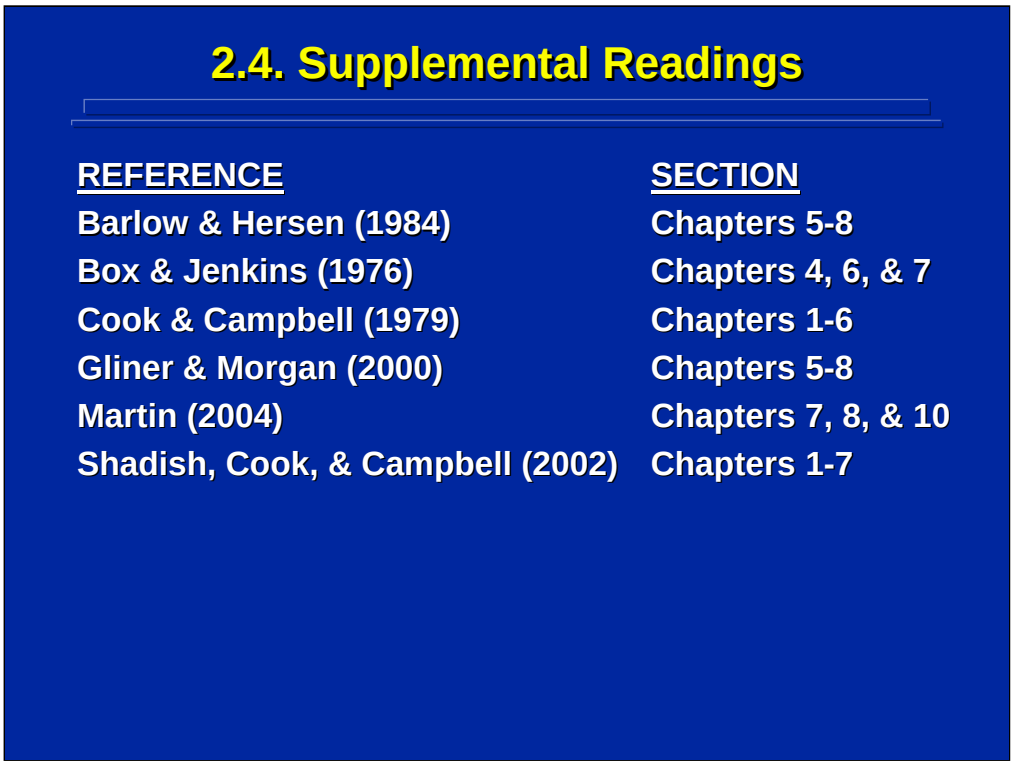
A way to control for variability among subjects is to have a homogeneous group of subjects. For example, balancing across experience level by experimental design can control for some of the subject variation. Randomly assigning subjects to treatment conditions is key to controlling subject variability and provides the major distinction between randomized and quasi-experimental designs.

The randomized experimental designs most often in human factors research are ANOVA designs. The reason that ANOVA is the mostly used is that human factors research deals with many factors that occur simultaneously, and each factor has more than one level. The structure of the ANOVA design allows the researcher the ability to test a family of hypotheses on one data set. Consequently, the emphasis of this reference material is focused on basic and advanced ANOVA designs.

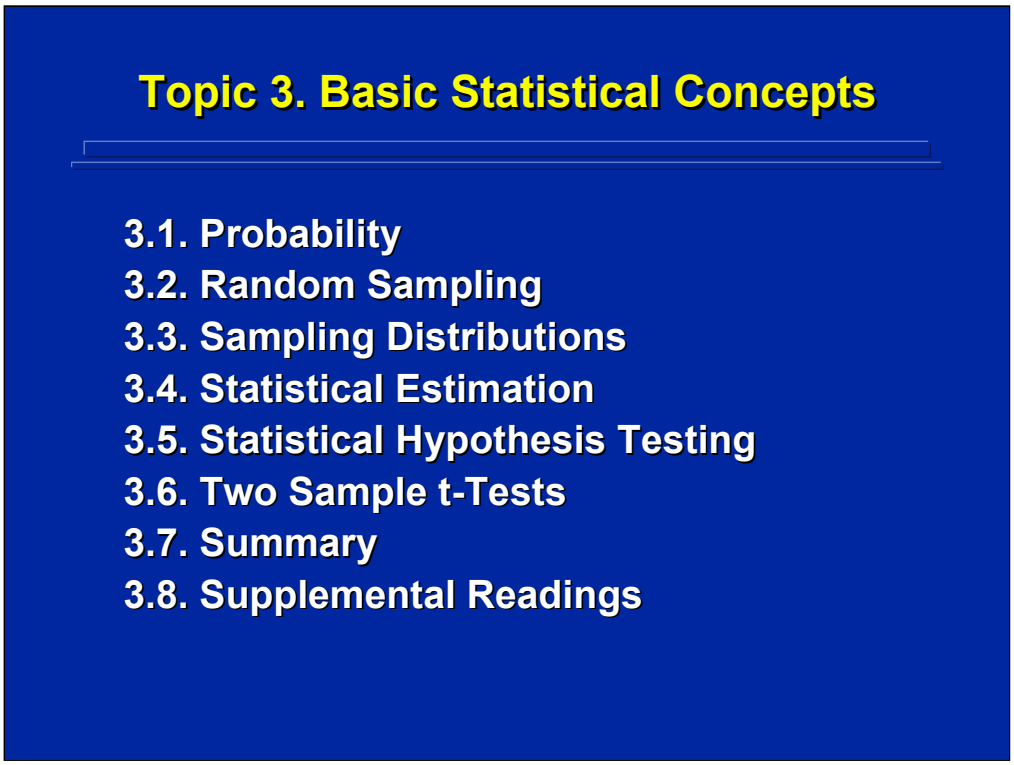



By way of summary, there are two categories of experimental designs: quasi-experimental and randomized experimental designs that can provide quantitative data for statistical analyses leading to casual inferences. The major distinction between these two categories of experimental design is the ability to use random assignment of subjects to treatment conditions. If random assignment is not possible, then the researcher can only use a quasi-experimental design.

In terms of overall implications, randomized experimental designs can provide controls for all four threats to validity discussed by Cook and Campbell (1979). In choosing an experimental design alternative, one should concentrate on internal and external threats that may be present. Randomized experimental designs provide the best control for internal and external threats to validity. Consequently, the focus of this reference material will be on randomized experimental designs with an emphasis on ANOVA.

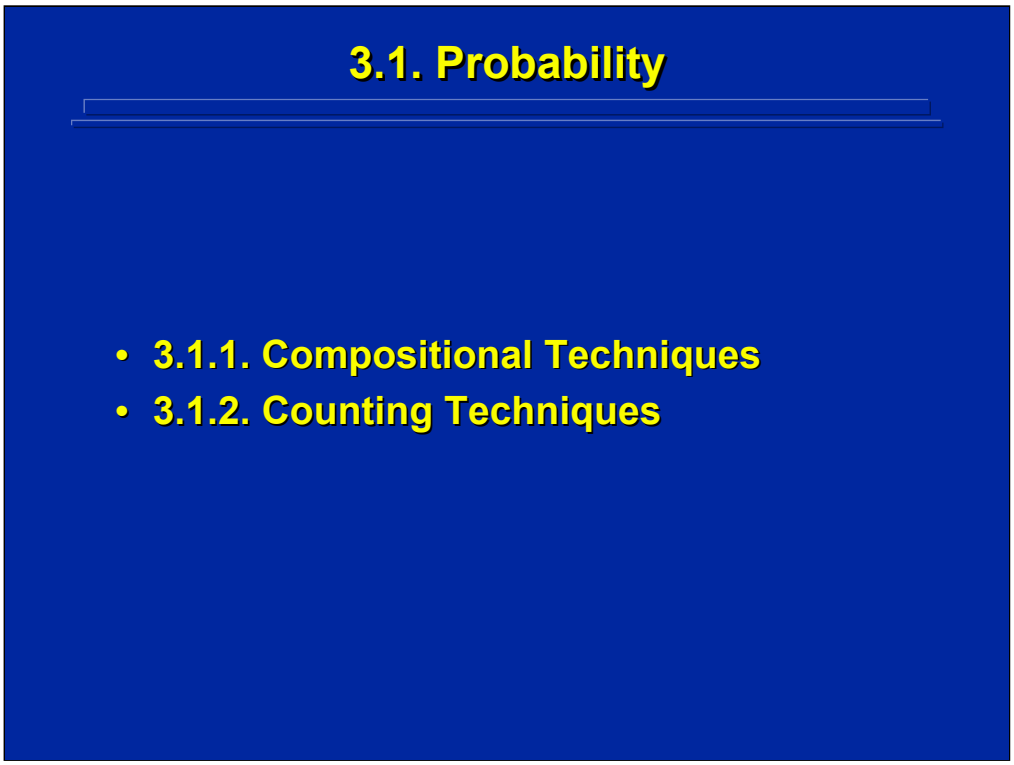



Barlow and Hersen (1984) provide an extensive discussion of alternative baseline experiments. Cook and Campbell (1979) provide the classic textbook on quasi-experimental designs as well as provide an overview of gain score and time series analysis. Shadish, Cook, and Campbell (2002) provide an update to the classic Cook and Campbell (1979) text. Box and Jenkins (1976) provide comprehensive time series analysis procedures. Finally, both Gliner and Morgan (2000) and Martin (2004) provide a general overview of quasi-experimental and randomized experimental designs.

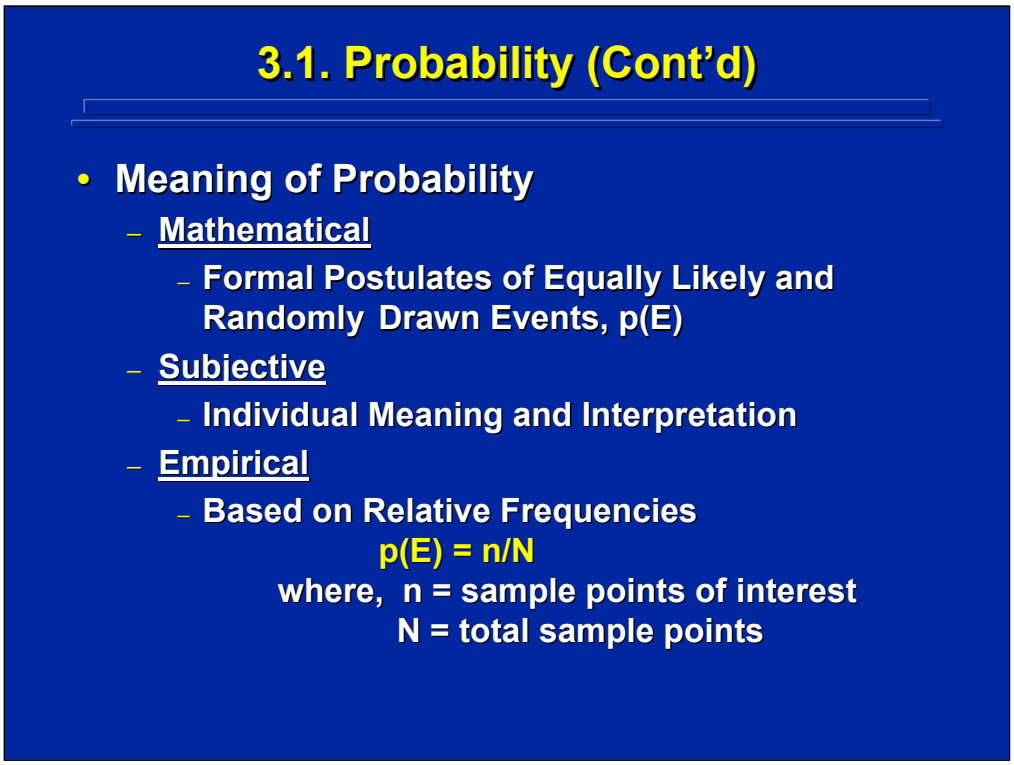



The purpose of this topic is to summarize basic statistical concepts that are fundamental to experimental design. This reference material is not a tutorial on basic statistics. Rather, the material highlights and summarizes key statistical concepts. It is assumed that users of this reference material already have a background in introductory descriptive and inferential statistics. Consequently, this topic is designed as a review of basic concepts without providing detailed descriptions or mathematical derivations. Users should refer to a textbook on introductory statistics for details on the concepts summarized in this topic if they are not familiar with them.

The concepts of sampling distributions, the F-distribution, and statistical hypothesis testing are critical to understanding the experimental design topics covered in this reference. The user should review these topics carefully as well as refer to the supplemental readings for additional details on various topics.




There are two techniques for determining mathematical and empirical values of probabilities, compositional and counting. The major components of each are reviewed since both are useful in experimental design.

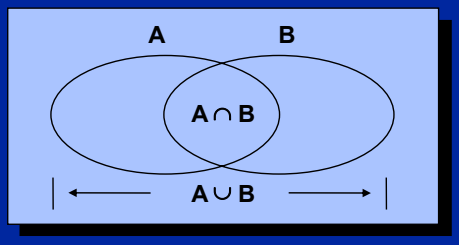



Probability is the basic structure of statistical analysis. There are three major definitions of probability, mathematical, subjective, and empirical. The mathematical definition provides formal postulates of equally likely and randomly drawn events, p(E). Subjective probability is the individual meaning and interpretation that one intuitively evaluates such as the probability of rain. Bayesian statistics, which are beyond the scope of this self-study material, deal with subjective probabilities mathematically.

In research design, the concentration is primarily on empirical probabilities and sample statistics. Empirical probability is based on relative frequencies, p(E) = n/N. The value of n equals total number of sample points; whereas N equals the total number of population points.

## 3.1.1. Compositional Techniques

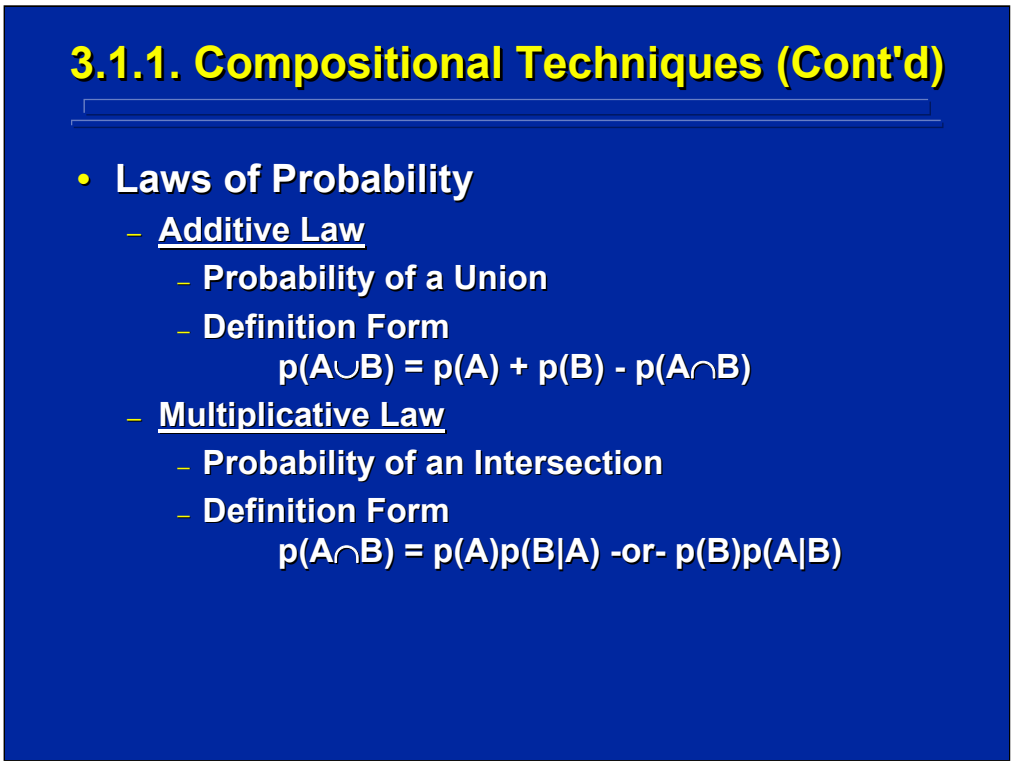
- Introduction
  - Set Theory - Compound Events
  - Laws of Probability
  - Relationships Used in Nonparametrics
- Compound Event Relationships



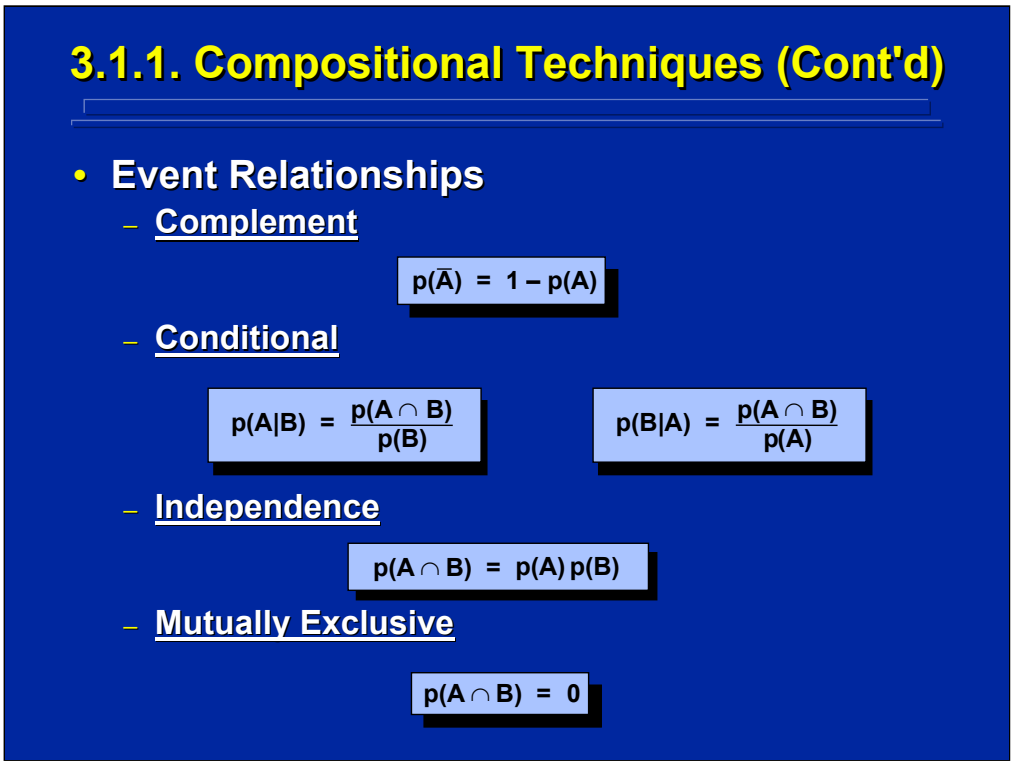

  - A = Event A
  - B = Event B
  - A∩B = Intersection = Joint Occurrence of Events A and B
  - A∪B = Union = Events A and B alone plus their Intersection

Compositional techniques are based on formal postulates of set theory that describe relationships of compound events that can be used to form the basic laws of probability. For example, the Venn diagram shown on this slide depicts the event space defined by event A and event B.

When more than one event is considered, relationships or compound operations between the events can be defined. Two fundamental relationships between events A and B are defined on this slide, the intersection and the union. The intersection is the overlap or joint occurrence of events A and B. The union includes all elements that belong to event A alone, event B alone, and the intersection of events A and B. Both the probability of an intersection and the probability of a union are important to experimental design. Several relationships based on these compound events are used in nonparametric statistical analyses.

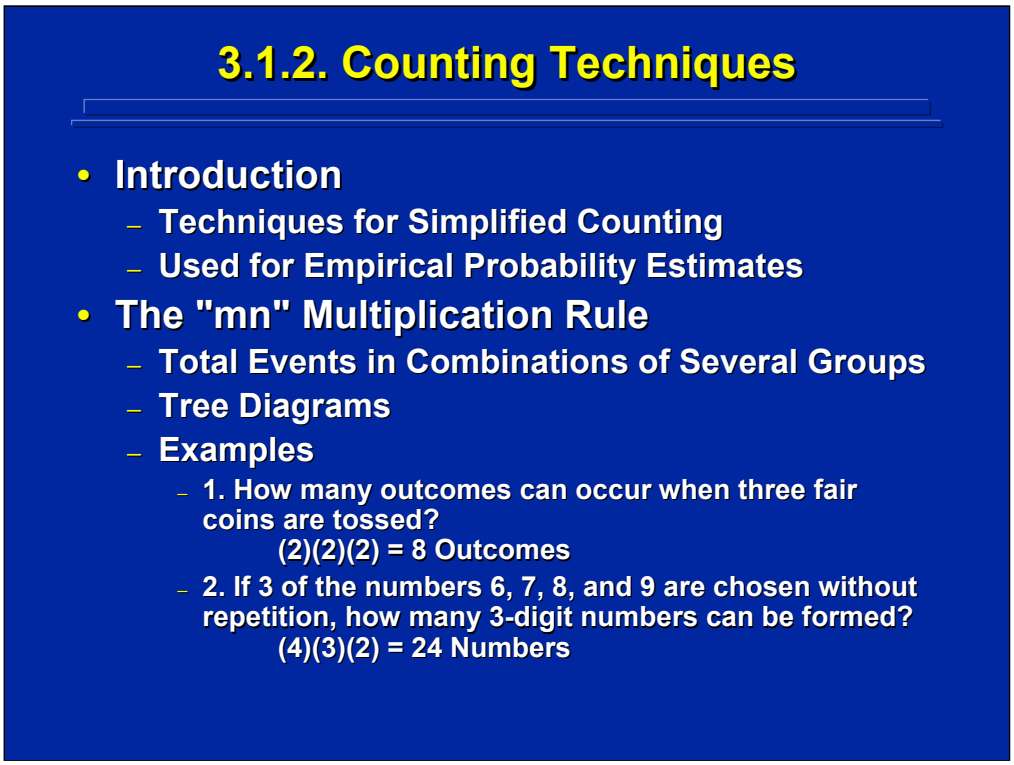



The mathematical laws of probability define the probabilities of the two basic relationships of compound events. The additive law defines the probability of a union which is equal to the probability of event A plus the probability of event B minus the probability of the intersection of events A and B. The multiplicative law defines the probability of an intersection, or joint probability of events A and B, as the probability of event A times the conditional probability of event A given event B occurs, or vice versa.




This slide defines four probability relationships based on operations of events A and B. Considering a single event A, one can define the complement of A as the probability of all elements or occurrences that are not event A.

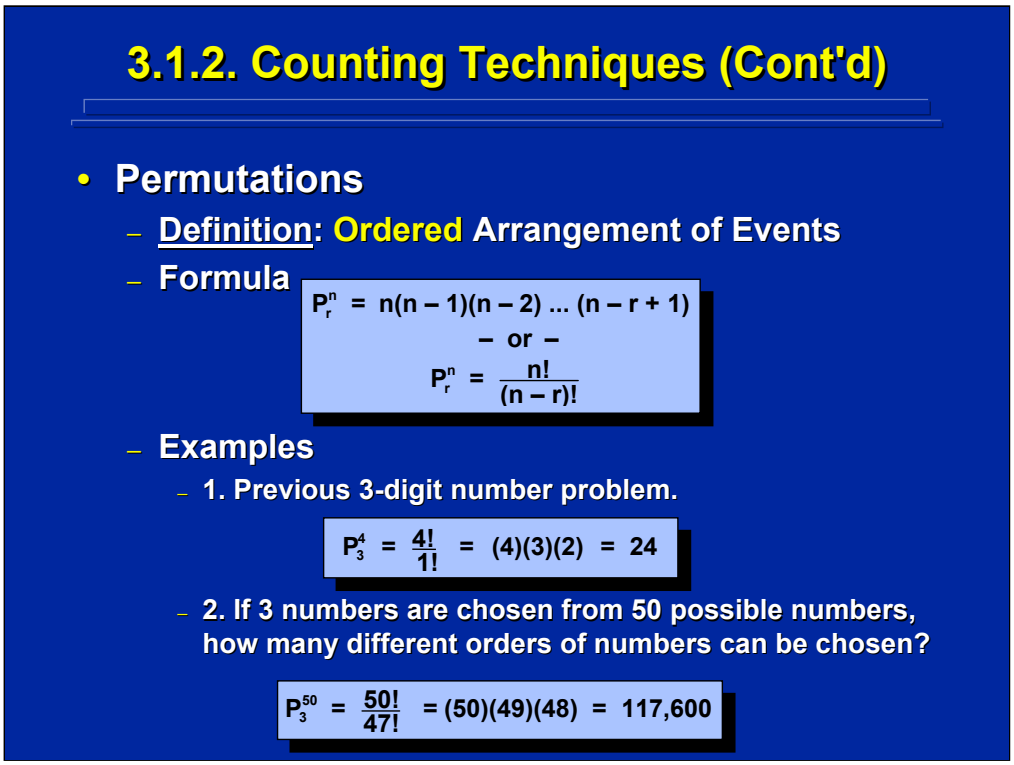
Considering two events A and B, one can define several relationships based upon the joint probabilities of events A and B. Conditional probability is the probability of event A occurring given that event B is present and vice versa. If events A and B are independent, the occurrence of event A does not depend upon the occurrence of event B. Independence can be defined as a joint probability equal to the probability of event A times the probability of event B. Finally, if events A and B are mutually exclusive, events A and B have no elements in common (i.e., no overlap exists in the Venn diagram of the compound event), and the joint probability of events A and B is zero.




To calculate probabilities, one needs to determine the number of possible distinctly different outcomes. Recall that p(E) = n/N where n is the total number of events of interest and N is the total number of all events. One could list every different outcome in simple cases to determine n and N, but this is not appropriate for large data spaces. Counting techniques employ rules or formulae for efficiently determining the frequency of outcomes.

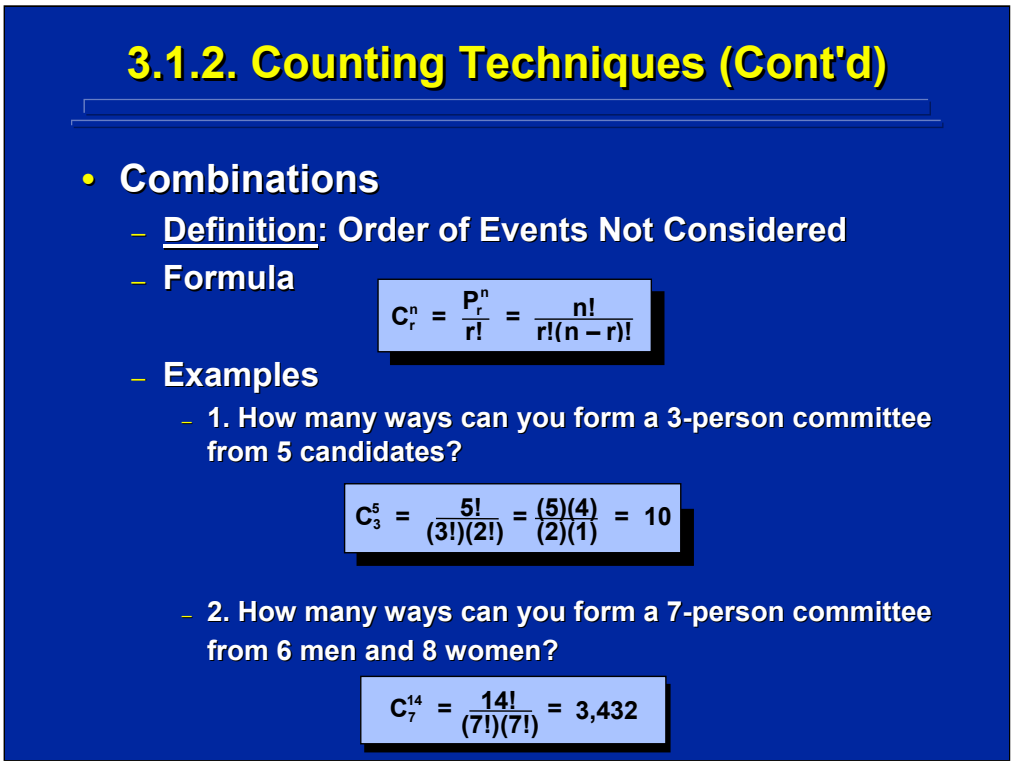
Although counting or tree diagrams can be used to determine the possible outcomes of a series of events, the “mn” rule can be used instead where “m” is the number of alternative outcomes for the first event and “n” is the number of outcomes for the second event in the series. The rule shows the multiplicative relationship between the number of alternative outcomes in each event. Obviously, the “mn” rule can be extended to a series of more than two events.

Two examples using the “mn” rule are shown on this slide each involving a series of three events. In the first example, each coin toss has 2 outcomes yielding 8 possible outcomes in a series of 3 tosses. In the second example, the solution is attained by determining how many ways there are to fill the “hundreds” position, then the “tens” position, and finally the “units” position in the resulting set of 3-digit numbers.



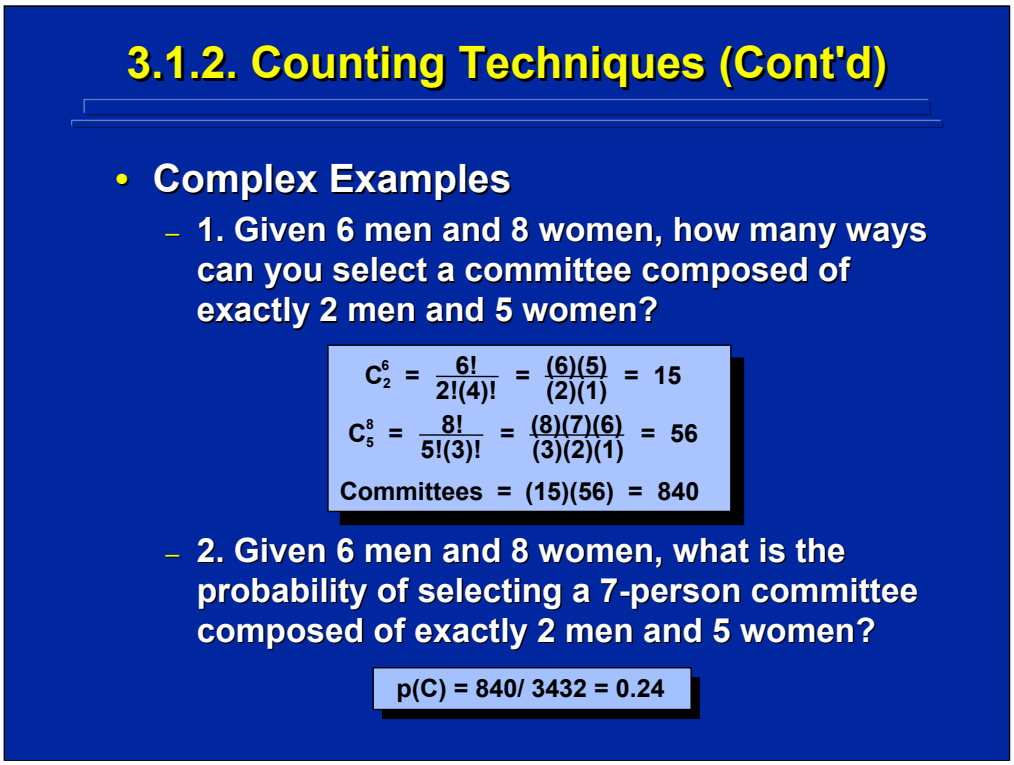

Permutations deal with ordered arrangements of events. The key word in a permutation is order. For example, in defining a two-digit number order is considered which means that the number 12 is different than 21. The general formula for calculating the number of permutations of “n” things taken “r” at a time is presented on this slide along with two examples.

First, consider that the 3-digit number example given on the previous slide to illustrate the “mn” rule is really a permutation since order is important in counting the number of alternatives in that particular example. The second example demonstrates the efficiency of using the permutation formula to determine the number of possible outcomes rather than listing all 117,600 possible outcomes.



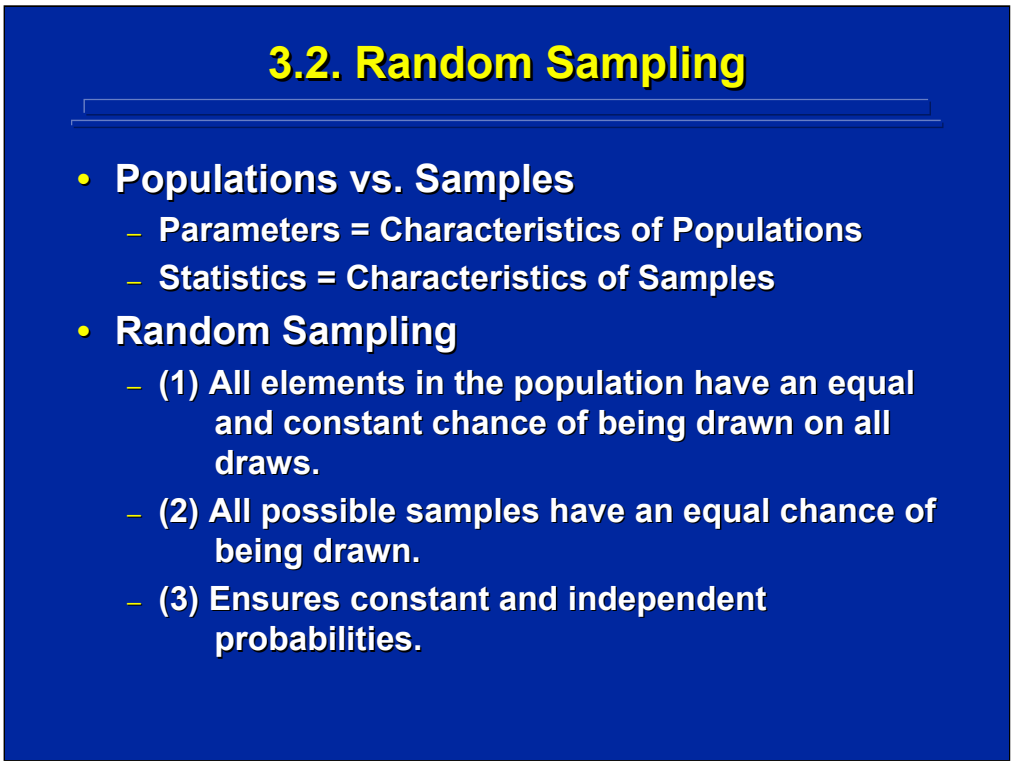

When order is not important the counting rule for combinations is used. If order is not important the combination of 12 and 21 are not unique outcomes since they each represent the same combination of the two digits, 1 and 2. So when “n” things are taken “r” at a time, there are fewer combinations than permutations. The number of combinations is equal to the number of permutations of “n” things taken “r” at a time divided by “r” factorial.

Two examples of using the combinations counting rule is provided on this slide. In the first example, the order of choosing a particular person for the committee is not important; one is counting the number of combinations, not permutations, of 5 people chosen 3 at a time. In the second example, one is not interested in how many men and women are chosen for the 7-person committee from the total of 14 eligible people. So, the outcome actually translates to the number of combinations of 14 people taken 7 at a time.



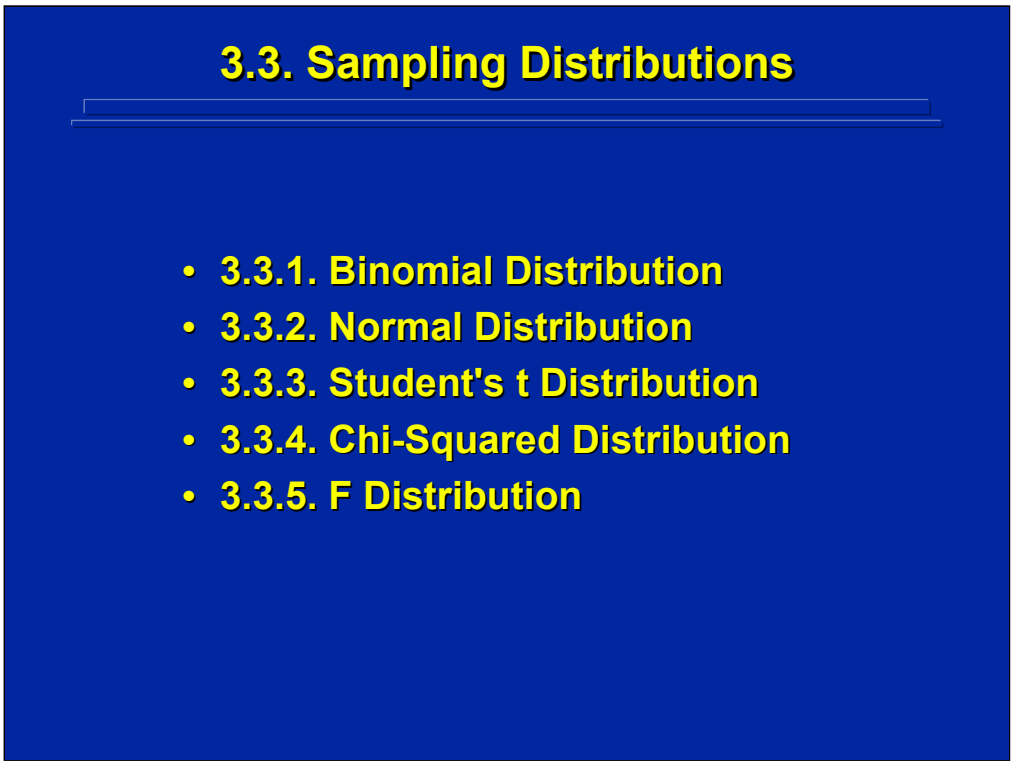

In more complex situations the counting rules can be used in combination. The first example on this slide combines two calculations for combinations and then the “mn” rule to obtain the total number of possible committees. First one determines the 15 combinations of men (choosing 2 of 6). Then one determines the 56 combinations of women (choosing 5 of 8). Finally, one determines through the “mn” rule that a total of 840 committees can be formed by the 15 combinations of men and 56 combinations of women. This is a good example of choosing the right set of counting techniques to determine the total number of alternatives without listing and counting all possible outcomes.

The second example on this slide demonstrates the use of counting techniques to determine an empirical probability. The probability of the committee, p(C), equals n/N. The value for n equals 840 committees composed of exactly 2 men and 5 women as calculated in the first example on this slide. The value for N equals all 3,432 possible 7-person committees as calculated in the second example of the previous slide. Consequently, the probability equals 0.24.




Sampling is a key component of experimental design. Data collected from samples are used to infer conclusions about populations. These samples are drawn randomly during data collection to avoid bias in the inferential process. Parameters are characteristics of populations. Greek letters will be used to list parameters. Statistics are characteristics of samples. Roman letters will be used to list statistics.

Three key characteristics of random samples are shown on this slide. For purposes of experimental design, random sampling of subjects assigned to treatment combinations has the restriction that an equal number of subjects will be assigned to each treatment condition in the experiment in order to keep sample size equal.




The probability distribution of a particular statistic is called the sampling distribution. By way of review, the following slides show general characteristics of the binomial, normal, student's t, chi-squared, and F distributions. The binomial is a discrete sampling distribution while the others are continuous. The sampling distribution used the most in human factors experiments is the F distribution, because it is used in ANOVA. Consequently, the F distribution will be described in more detail when discussing ANOVA.

## 3.3. Sampling Distributions (Cont'd)

- **Definition:**
  - A sampling distribution is a probability distribution that represents the likelihood of all the various values of a particular statistic for a particular sample size, n.
  - Mathematical Description - Appendix A in Winer, et al. (1991)
- **Common Sampling Distributions**
  - Binomial Distribution
  - Normal Distribution
  - Student's t Distribution
  - Chi-Squared Distribution
  - F Distribution

The distribution of values of a statistic calculated from samples is characterized by a sampling distribution. A sampling distribution is a probability distribution that represents the likelihood of all the various values of a particular statistic for a particular sample size, n. There are three critical elements in this definition. First, a sampling distribution is a cumulative probability density function, f (X). The area under this describing function sums to 1.0. Consequently, the probability of any particular value of X can be determined by integrating the area under the curve of a continuous probability density function. Second, a sampling distribution is unique for a particular statistic, which means that one would have separate sampling distributions for the mean, standard deviation, or variance. Most often a sampling distribution for means is used in experimental design. The third characteristic is that a sampling distribution is based on a particular sample size. Depending on sample size, n, the shape of the sampling distribution may vary significantly. One must know how sample size affects the sampling distribution.

The five sampling distributions listed on the bottom of this slide are most often used in experimental design. The binomial is a discrete probability distribution; whereas, the other four are continuous.

## 3.3. Sampling Distributions (Cont'd)

- **Example: Using a set of the five values, 4, 5, 6, 7, and 8, show the sampling distribution of the mean for a sample size of 3. (Heuckeroth, 2004)**

| All Possible Samples (n = 3) | Sample Means | Probability Distribution |
|---|---|---|
| 4 + 5 + 6 | 15/3 = 5.0 | 5.0 = 0.10 |
| 4 + 5 + 7 | 16/3 = 5.3 | 5.3 = 0.10 |
| 4 + 5 + 8 | 17/3 = 5.6 | 5.6 = 0.20 |
| 4 + 6 + 7 | 17/3 = 5.6 | |
| 4 + 6 + 8 | 18/3 = 6.0 | 6.0 = 0.20 |
| 5 + 6 + 7 | 18/3 = 6.0 | |
| 4 + 7 + 8 | 19/3 = 6.3 | 6.3 = 0.20 |
| 5 + 6 + 8 | 19/3 = 6.3 | |
| 5 + 7 + 8 | 20/3 = 6.6 | 6.6 = 0.10 |
| 6 + 7 + 8 | 21/3 = 7.0 | 7.0 = 0.10 |

Consider the simple example of constructing a sampling distribution as provided by Heuckeroth (2004) which aptly demonstrates the three critical elements involved. First, one must define all possible samples of 3 values that can be drawn from the five values. This is simply the number of combinations of 5 things taken 3 at a time or 10 possible samples as shown in the left column on this slide. Second, the 10 samples have a total of only 7 different means as shown in the center column. Finally, the probability of obtaining each of the 7 different mean values is calculated as shown in the right column. Plotting the values of each mean against the probability of obtaining that mean is the resulting sampling distribution. Obviously, this sampling distribution would be different if the number of possible values, the sample size, or the statistic calculated changed. Conceptually, however, every sampling distribution whether discrete or continuous has these three critical elements.

The five most common sampling distributions used in experimental design include the binomial, normal, student's t, chi-square, and F distribution. Each is reviewed separately. Conover (1999) provides a discussion of the discrete binomial sampling distribution in Chapter 1 and its various uses in Chapter 3. Winer, Brown, and Michels (1991) provide an overview of the continuous probability distributions in Chapter 2 as well as mathematical descriptions in Appendix A.

## 3.3.1. Binomial Distribution

- Discrete Probability Distribution
- Binomial Experiment
  - Based on Two Possible Mutually Exclusive Outcomes (e.g., "Success" and "Failure") where
    - p = Probability of Outcome 1
    - q = Probability of Outcome 2 = (1 - p)
    - n = Number of Independent Trials
    - k = Number of "Outcome 1" in "n" Trials
- Binomial Theorem, $(p + q)^n$

$$(p + q)^n = p^n + np^{n-1}q + \frac{n(n-1)}{2!}p^{n-2}q^2 + \ldots + q^n$$

Where the probability of any term is defined by:

$$C_k^n p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

The binomial distribution is a discrete probability distribution. Many nonparametric statistical analyses use the binominal distribution during hypothesis tests that evaluate frequencies of discrete categories. The binomial distribution has two possible and mutually exclusive outcomes often known as “success” and “failure”. In a binominal experiment of “n” independent trials, “k” is defined as the number of successes. The probability of various outcomes in a binominal experiment is defined by the binomial theorem as shown on this slide where each term, or possible outcome, is a combination of “n“ things taken “k” at a time.

## 3.3.1. Binomial Distribution (Cont'd)

- **Example: Find the probability of 2 heads in 3 tosses of a fair coin.**

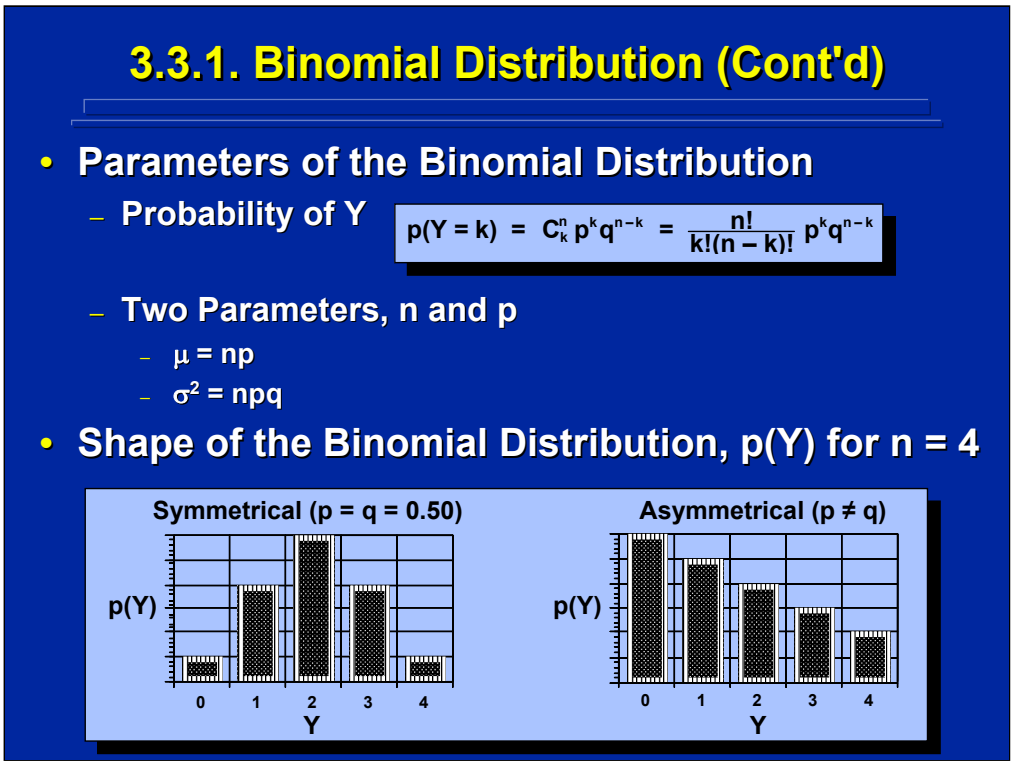
| Outcome | Probability | Y |
|---|---|---|
| H,H,H | $p^3$ | 3 |
| H,H,T | $p^2q$ | 2 |
| H,T,H | $p^2q$ | 2 |
| H,T,T | $pq^2$ | 1 |
| T,H,H, | $p^2q$ | 2 |
| T,H,T | $pq^2$ | 1 |
| T,T,H | $pq^2$ | 1 |
| T,T,T | $q^3$ | 0 |

| Y=k | p(Y=k) | | |
|---|---|---|---|
| 3 | $p^3$ | $= (1/2)^3$ | $= 1/8$ |
| 2 | $3p^2q$ | $= 3p^2(1/2)$ | $= 3/8$ |
| 1 | $3pq^2$ | $= 3(1/2)(1/2)^2$ | $= 3/8$ |
| 0 | $q^3$ | $= (1/2)^3$ | $= 1/8$ |

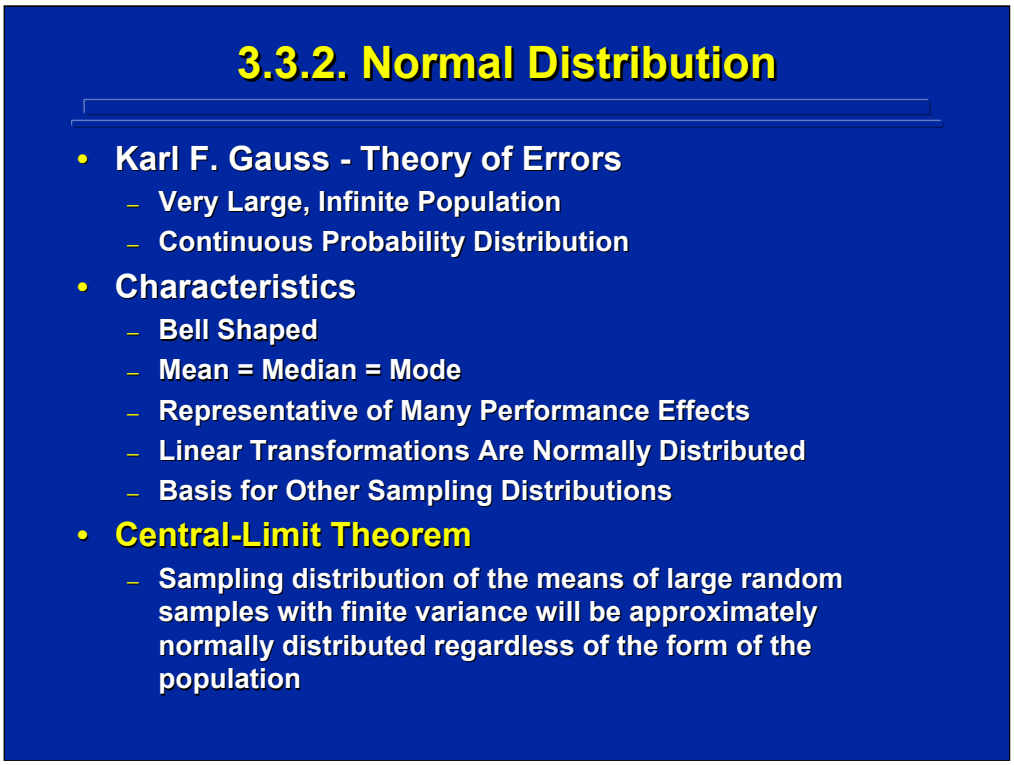
$$C_k^n p^k q^{n-k} = \frac{3!}{2!(3-2)!} p^2q = 3(1/2)^2(1/2) = 3/8$$

Consider a simplistic binomial experiment consisting of 3 tosses of a fair coin where heads is considered “success” and tails is considered “failure”. The probability of getting 3 heads, 2 heads, 1 head, or 0 heads is simply the relative frequency of each alternative divided by 8, the total number of possible outcomes. Alternatively, the probability can be calculated directly using the binomial theorem. As shown on the bottom of this slide, the binomial theorem is used to determine the probability of obtaining 2 heads in 3 tosses. Note this is a question of combinations, because the order in which 2 heads occurs in the 3 tosses is irrelevant.

Obviously, as the total number of tosses increases the determination of probabilities in this binomial experiment is much easier to compute using the binomial theorem formula than by directly listing and counting all possible outcomes. Consider, for example, counting all possible outcomes of 100 coin tosses instead of just the 3 coin tosses shown on this slide. However, by enumerating and counting all outcomes in this small (i.e., n = 3) binomial experiment, the validity of the computational formula based on the binomial theorem is readily apparent.

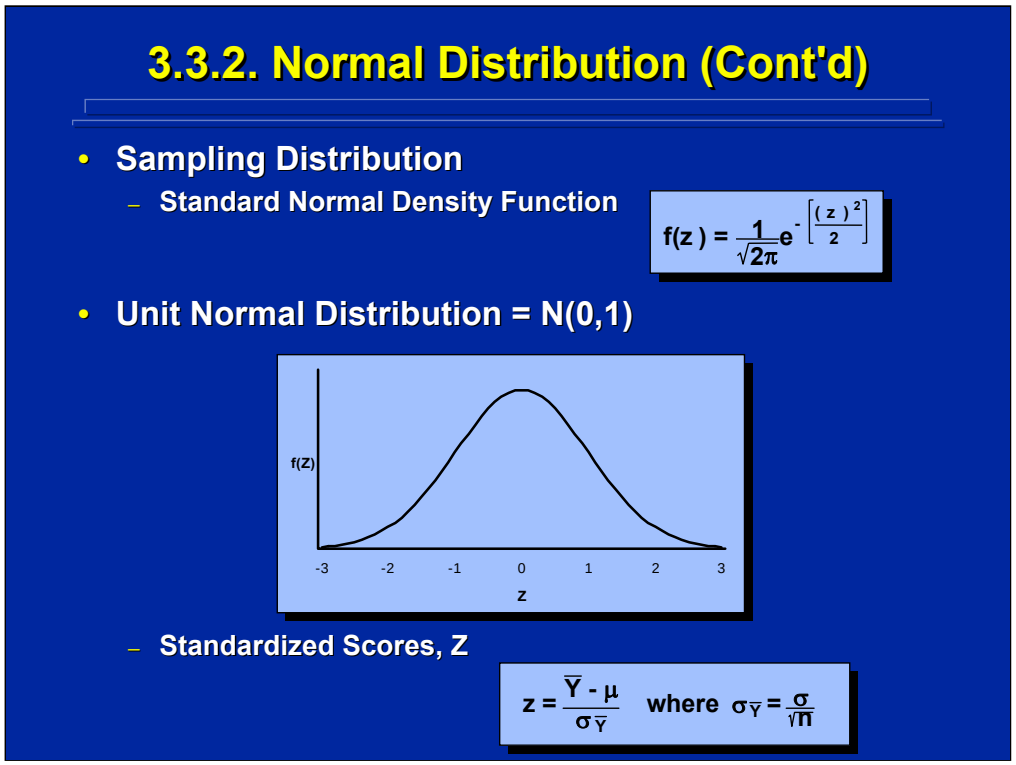



The binomial distribution is defined as the probability of Y, success, in terms of two parameters, n and p. These parameters can be used to find the mean and variance. If p and q (i.e. 1-p) are equal, then there is a symmetric distribution. If p and q are not equal, there is an asymmetrical distribution. The direction of asymmetry depends on whether p or q is larger. In the asymmetrical example shown on the slide for 4 trials, the probability of success, p, is smaller than failure, q.



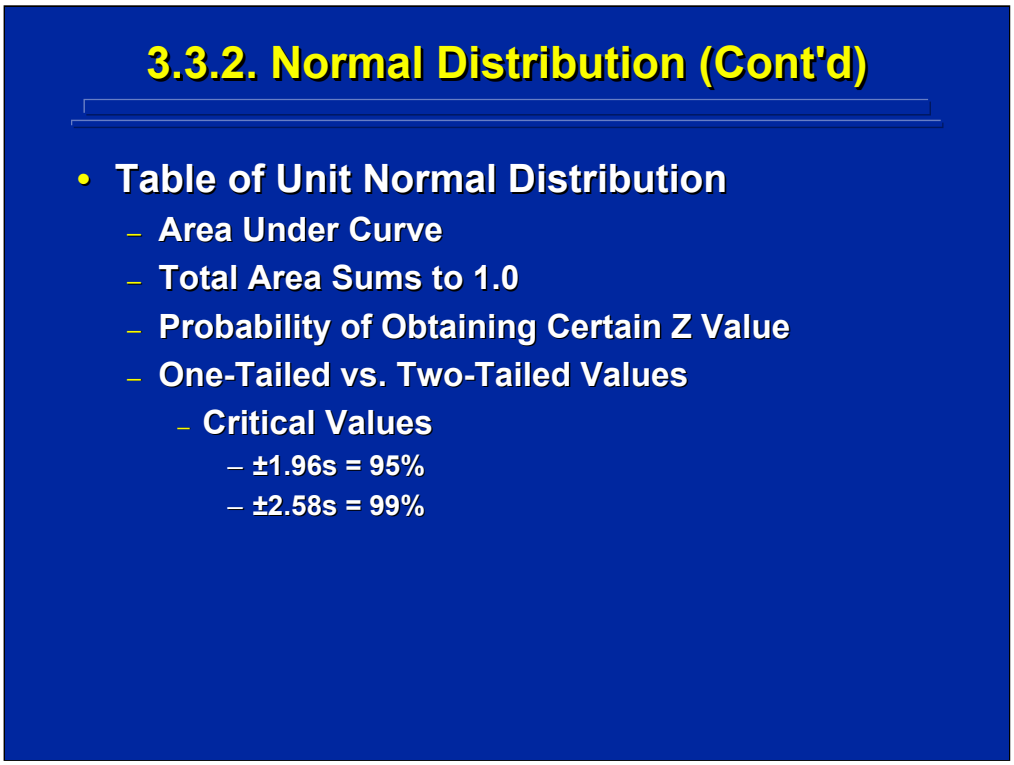

The most common continuous probability sampling distribution is the normal distribution. It was defined by Karl Gauss, while studying the theory of errors. The bell or symmetrical shape implies that the three descriptive statistics of central tendency, mean, median, and mode, are all equal. The normal distribution is frequently used because it is representative of many performance effects. Linear transformations of any set of normally distributed scores are also normally distributed and will not change the shape of the distribution. The normal distribution also becomes the basis for other sampling distributions used in experimental design that assume a normal distribution of variables.

One important aspect of the normal distribution is the central limit theorem. This theorem states that the sampling distribution of means of large random samples with finite variance will be approximately normally distributed regardless of the form of the underlying population. Consequently, the normal distribution becomes a robust sampling distribution for inferential statistical comparisons on a wide variety of data collected in experiments.



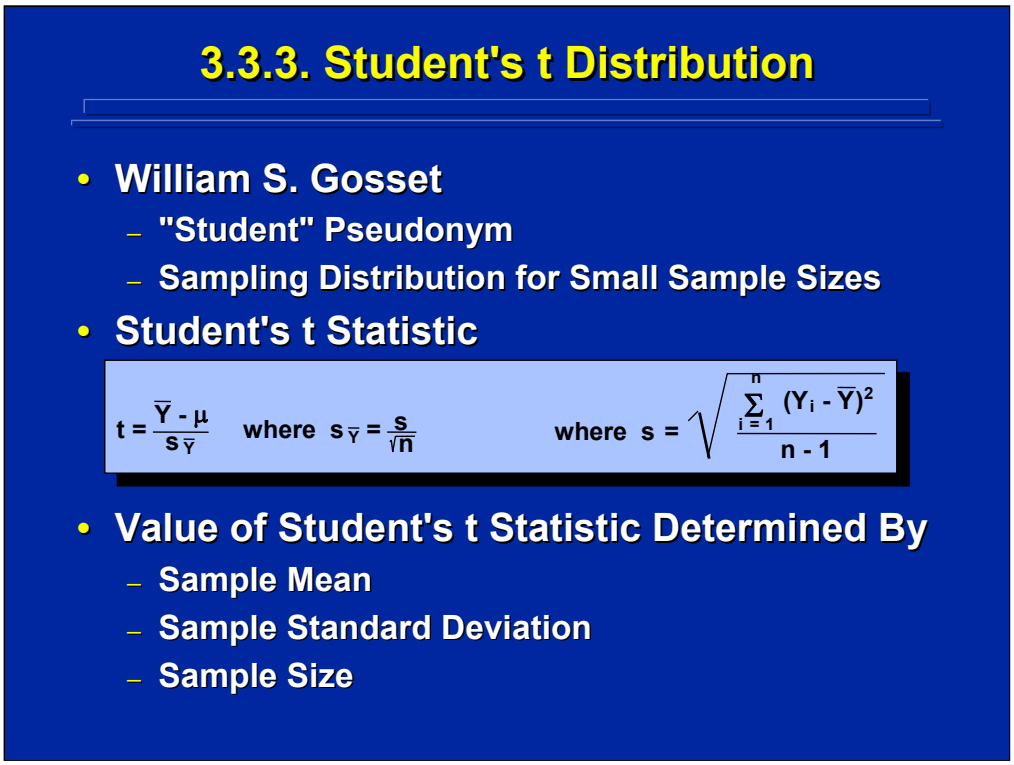

The probability density function for the normal distribution is defined by the equation shown on this slide. The three parameters of the normal sampling distribution, N($\mu$, $\sigma$), are sample size, N, population mean, $\mu$, and population standard deviation, $\sigma$. The shape of the normal distribution is always symmetrical, but the peakedness depends on the population standard deviation. As the population standard deviation decreases the normal distribution becomes more leptokurtic or peaked. As the population standard deviation increases, the normal distribution becomes more platykurtic or flat.

A special form of the normal distribution is the unit normal distribution which is based on standardized scores, Z scores. The Z scores are calculated according to the formula shown on the slide. The Z scores have a mean of 0 and a standard deviation of 1, thereby providing the designation parameters of N(0,1). The unit normal distribution is commonly referred to as the Z distribution, and its shape is shown on this slide.




The table of the unit normal distribution states the area under the curve at various standard deviations. Since this is a probability density function, the total area sums to 1. Two critical values used in hypothesis testing are the number of standard deviations of the unit normal distribution that contain 95% (i.e. plus and minus 1.96 standard deviations) and 99% (i.e. plus and minus 2.58 standard deviations) of the area.

One-tailed tests have all the remaining area of the distribution under one end of the distribution; whereas two-tailed tests have the remaining area equally divided at both ends of the distribution. This slide shows the two-tailed value of the unit normal distribution at the 95% and 99% confidence level. A one-tailed test which assumes a difference in only one direction is less conservative (i.e. easier to obtain significance), because the Z value would be smaller than the two-tailed test at any confidence level. Typically the more conservative two-tailed tests are used in experimental design for hypothesis testing, because the difference obtained from the data can be either greater or smaller than the expected value .




A variant of the normal distribution is the Student's t distribution. William Gosset developed this distribution under the pseudonym of Student. The student's t is a sampling distribution based on small sample sizes. The difference between the t-distribution and the normal distribution is that the standard error is based on the sample standard deviation not the population standard deviation. The student's t statistic is determined by the sample mean, sample standard deviation, and the sample size.

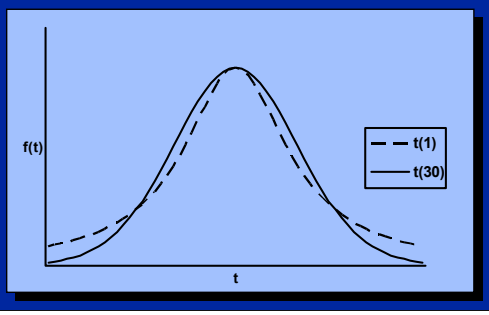
## 3.3.3. Student's t Distribution (Cont'd)

- **Probability Density Function**

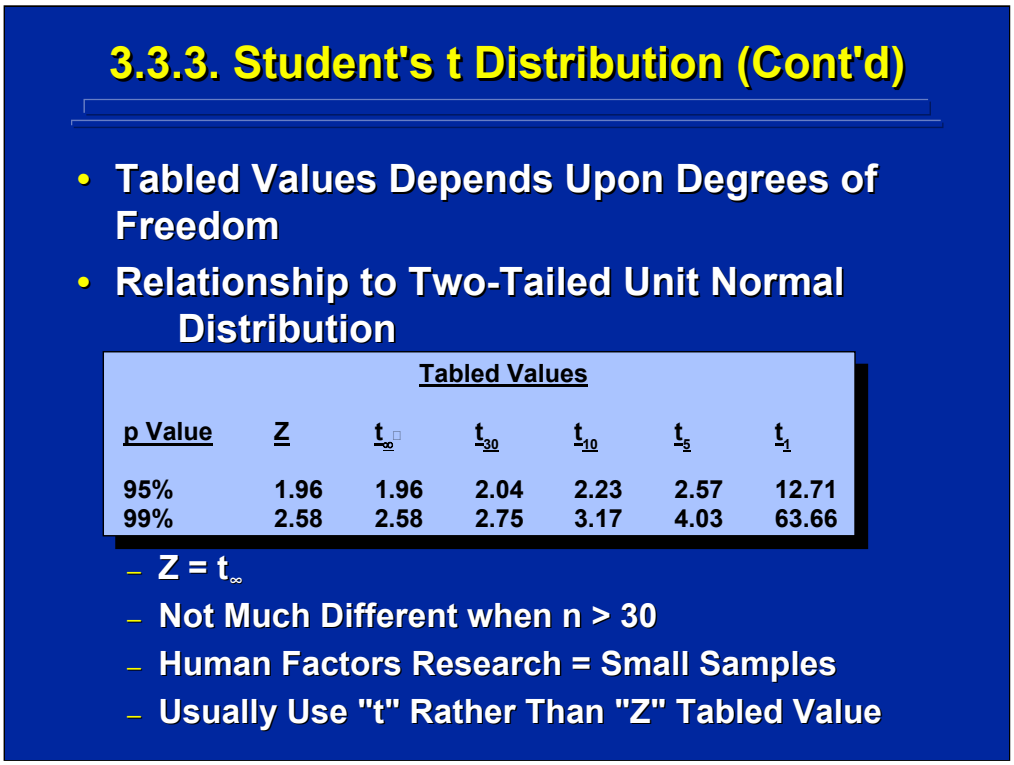
$$f(X) = \frac{c}{\left(1 + \frac{t^2}{\nu}\right)^{\frac{(\nu+1)}{2}}}$$

where, $c$ = Constant
$\nu$ = Degrees of Freedom

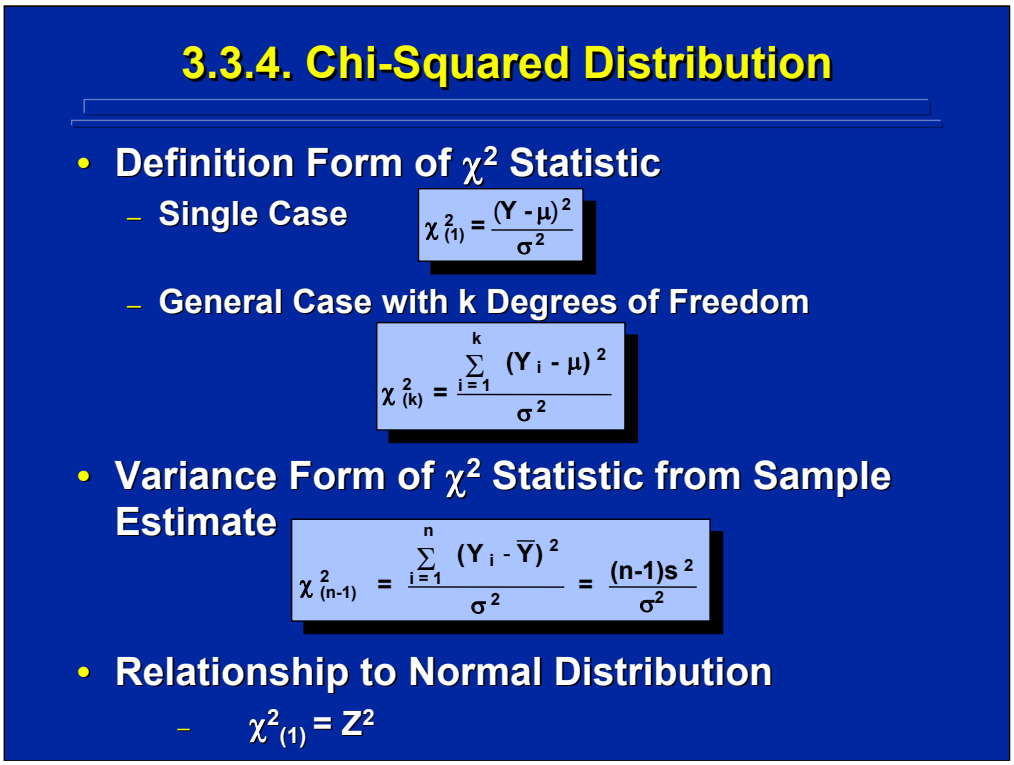
- **Sampling Distribution**




The formula for the probability density function as well as the shape of the sampling distribution is shown in this slide. The shape of the t distribution is always symmetrical and is more leptokurtic than the normal distribution until sample size becomes large.




This slide compares critical values of the t distribution with the unit normal, Z, distribution. The tabled values depend upon degrees of freedom. As sample size, n, increases, the t distribution approaches the normal distribution. Once sample size, n, gets above 30, there is only a slight difference between the t and Z values. Because human factors research usually deals with small sample sizes, the t rather than the Z distribution is used primarily. Again this is the conservative approach, because the t-tabled value is larger than the Z-tabled value.




Another sampling distribution often used in experimental design is the chi-squared distribution. The chi-squared statistic is defined as the squared difference of the observed value, Y, from the population mean, $\mu$, divided by the population variance, $\sigma^2$.

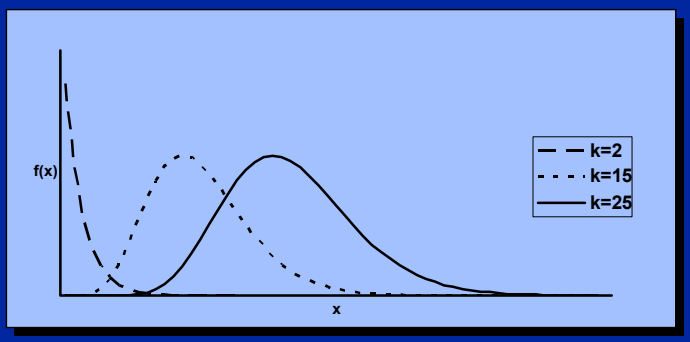
This slide shows the definitional formula for both a 1 degree of freedom chi-square and a k degree of freedom chi-square statistic. The variance form of the chi-squared statistic is important because it is used in defining the F statistic in ANOVA. Finally, a chi-square statistic with 1 degree of freedom is equal to the unit normal statistic squared.

## 3.3.4. Chi-Squared Distribution (Cont'd)

- **Probability Density Function**

$$f(X) = c(\chi^2)^{\sqrt{\nu-2}} e^{-\sqrt{\chi^2}}$$

- **Sampling Distribution**




The formula for the chi-square probability density function is shown on this slide along with the shape of the resulting sampling distribution. Note the shape changes depending on the number of degrees of freedom, $\nu$, of the chi-square statistic. The chi-squared sampling distribution is positively skewed (i.e., tailing to the right rather than the left,) and it becomes more symmetrical as the degrees of freedom increase.

## 3.3.5. F Distribution

- Sir Ronald Fisher
- Definition of F: Ratio of two independent $\chi^2$ variables each divided by their appropriate df.

$$F = \frac{\chi^2_{(\nu_1)} / \nu_1}{\chi^2_{(\nu_2)} / \nu_2}$$

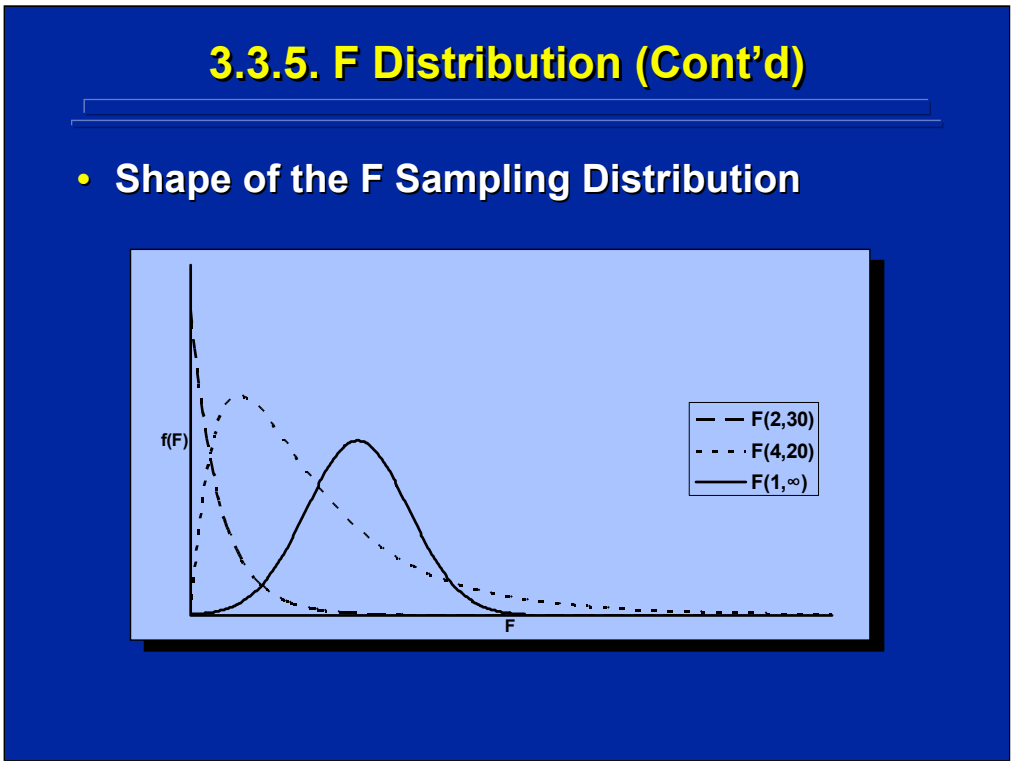
- Two independent sample variances where $s_1^2 \geq s_2^2$

$$F = \frac{s_1^2}{s_2^2} \quad \text{where } s^2 = \frac{\sigma^2 \chi^2_{(\nu)}}{\nu}$$

- Assumptions
  - Parent populations are normal.
  - Samples are drawn independently.
  - Population variances are equal.

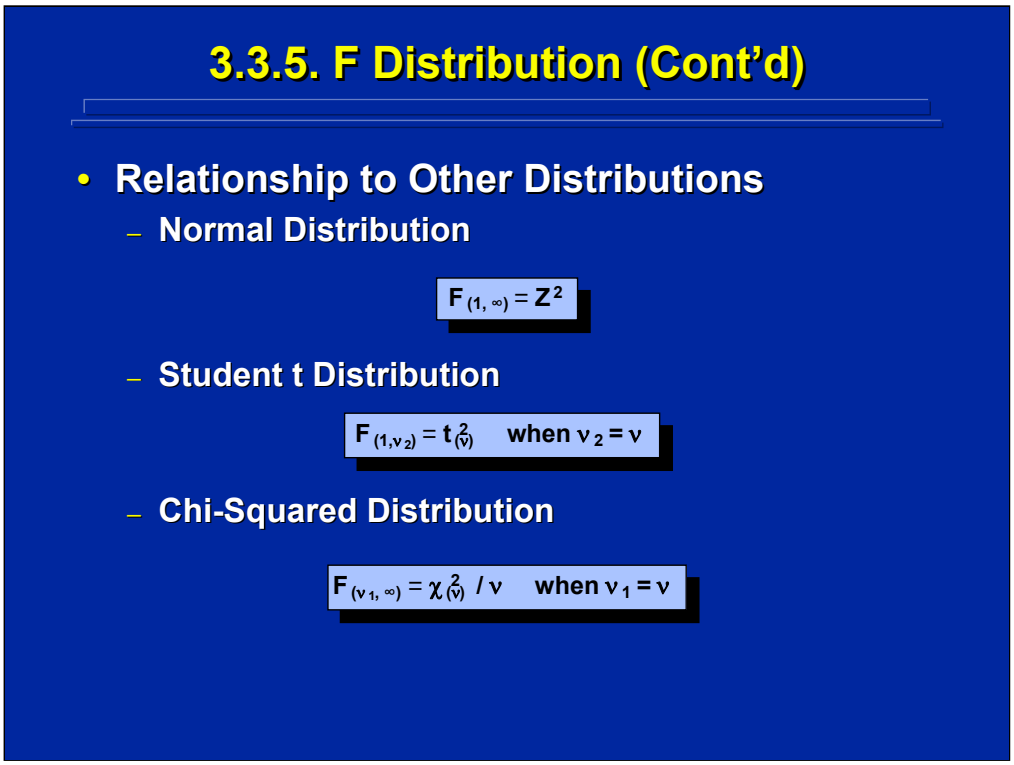
The F-ratio was first derived and developed by Sir Ronald Fisher. The F statistic is defined as the ratio of two independent chi-square variables each divided by their appropriate degrees of freedom as shown on the slide. The F statistic is used in ANOVA experimental designs and is estimated by two sample variances. Looking at the variance form of a chi-square statistic, the ratio of two sample variances forms a legitimate F-ratio if the population variance, $\sigma$, for both estimates is equal.

There are three major assumptions of an F distribution. The parent populations are assumed normal because, by definition, the F-ratio is a ratio of two chi-square variables that are drawn from a normal population. The assumption that the samples are drawn independently also comes from the definition of a chi-square statistic. The assumption that two samples have equal population variance is based on the variance formula for a chi-square statistic and results in the ratio of two sample variances as shown on this slide. Consequently, the F statistic can be simply stated as the ratio of two independent sample variances.

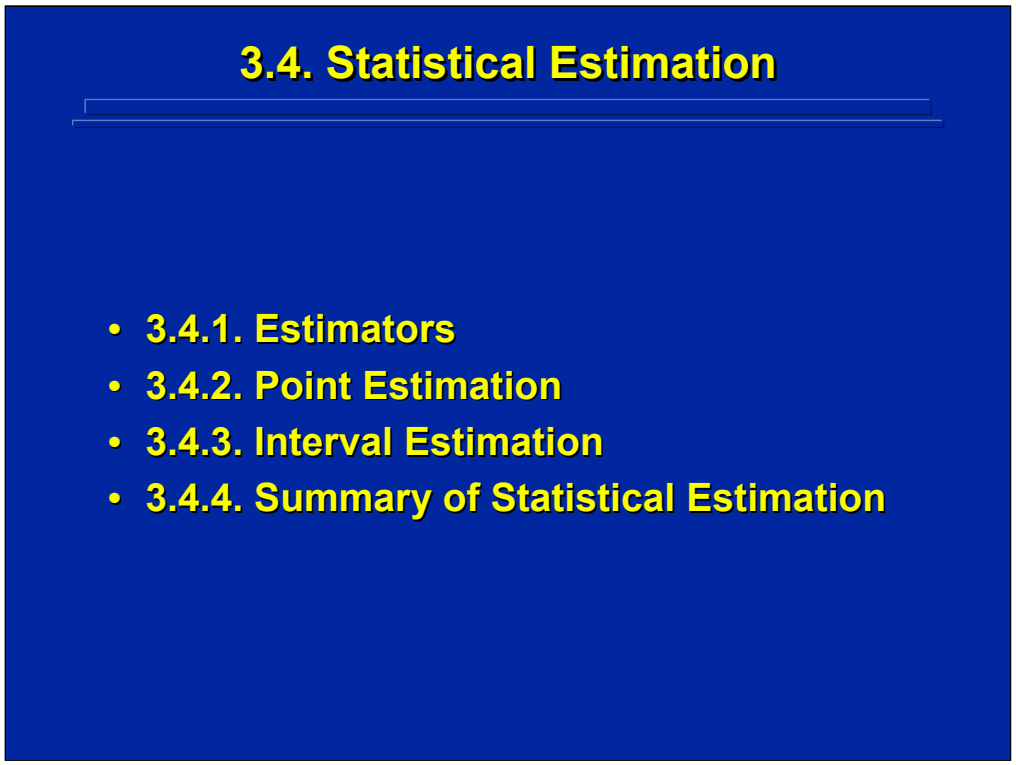



Any F statistic has two sets of degrees of freedom associated with it, the degrees of freedom for the variance in the numerator, and the degrees of freedom for the variance in the denominator. The shape of the sampling distribution of the F statistic is really a family of distributions that depend upon the degrees of freedom of the numerator and denominator of the F-ratio.

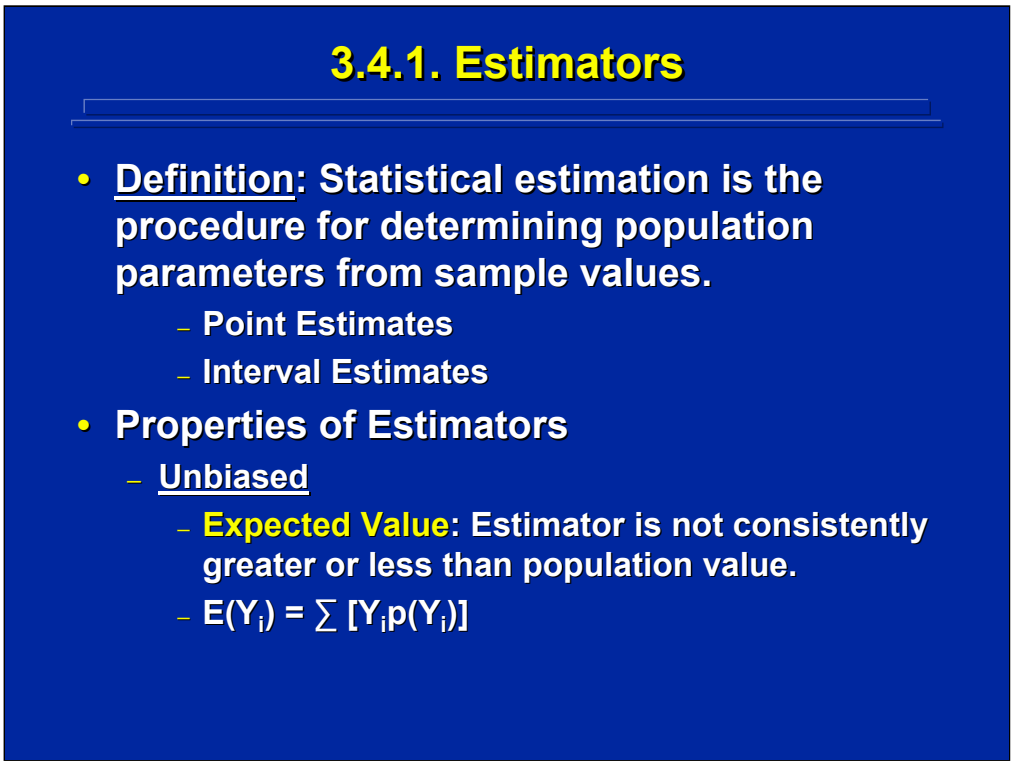
The three stylized shapes of the F sampling distribution shown on this slide for various degrees of freedom on the numerator and denominator illustrate that the F sampling distribution is highly positively skewed when only a few degrees of freedom exist. As the degrees of freedom increase, the F distribution becomes bell shaped like the unit normal distribution.




Since the F statistic is a ratio of two independent chi square statistics and the sampling distribution starts out highly positively skewed and approaches the normal distribution as degrees of freedom increase, the F sampling distribution is directly related to the chi square, student's t, and unit normal distribution. There is a direct relationship of the values in the F table with the tabled values of the normal, student's t, and chi-squared distributions as shown in the formulae listed on this slide.

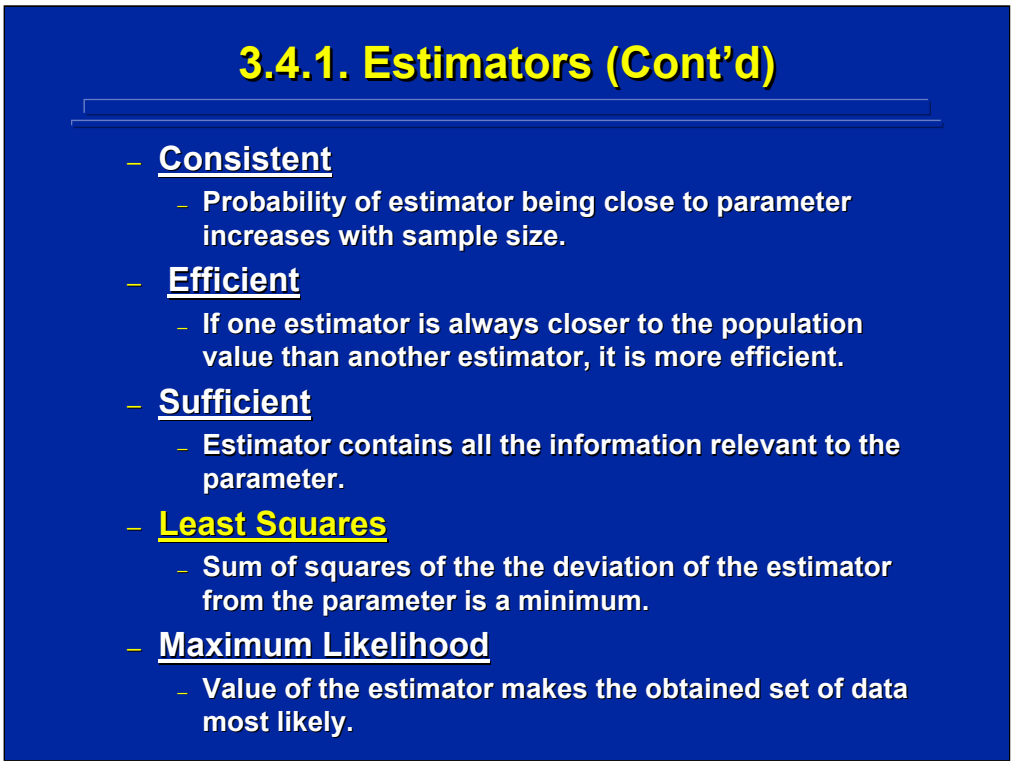



A major component of inferential statistics is the estimation of population parameters (i.e., means and variances) from sample statistics. In this section, the general characteristics of good estimators, the calculation of various point estimates useful in experimental design, and interval estimation calculation are reviewed.

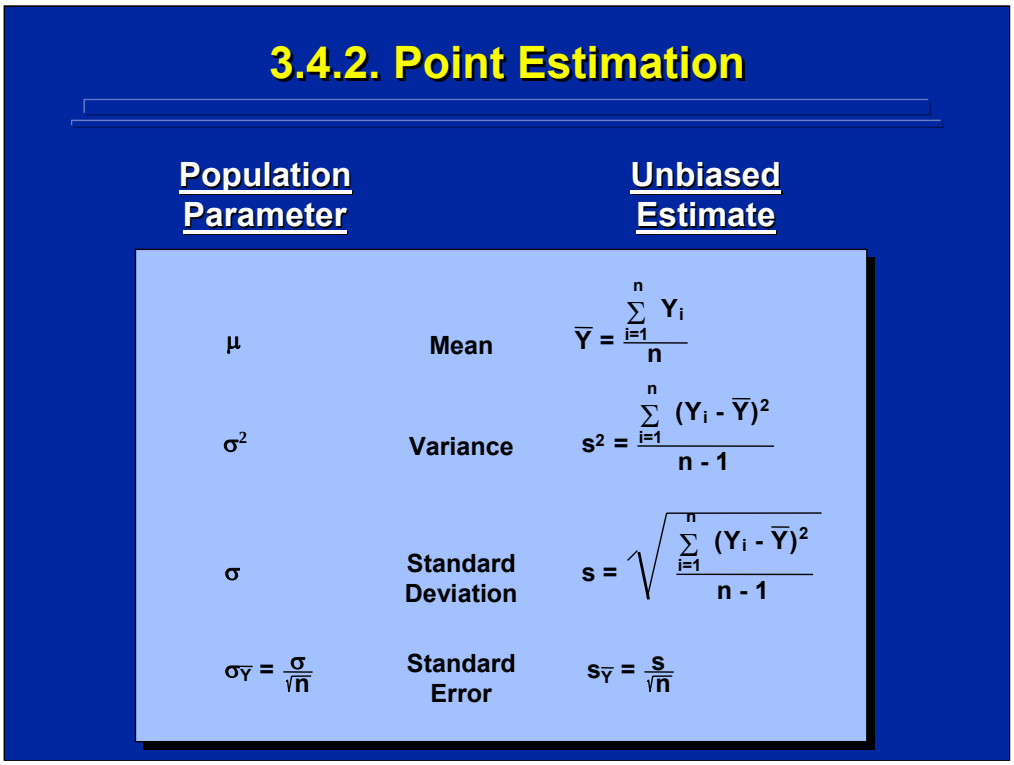



Statistical estimation is the procedure for determining population parameters from sample values. In other words, researchers try to estimate a certain numerical characteristic of the population of interest in their research. There are two ways to do this. A point estimate is a single number. An interval estimate is a probability statement that the point estimate will fall somewhere between a specified upper and lower limit. Consequently, one can either provide one number or a range of numbers when calculating a parameter estimate.

There are several mathematical properties of estimators that can be used to determine the “goodness” of the estimator. The first is an unbiased estimator, which means that the estimator is not consistently greater or less than the population value. The expected value of a statistic, as defined by the formula on the slide, is an unbiased estimate of the population value.




Other properties of good estimators include consistent, efficient, sufficient, least squares, and maximum likelihood. The definitions of each of these properties are listed on the slide. The least squares property is important in experimental design. A least squares criterion means that the sum of squares of the deviation of an estimator from the parameter is a minimum. Of all the common mathematical properties of an estimator, unbiased and least squares properties are of major concern in experimental design.



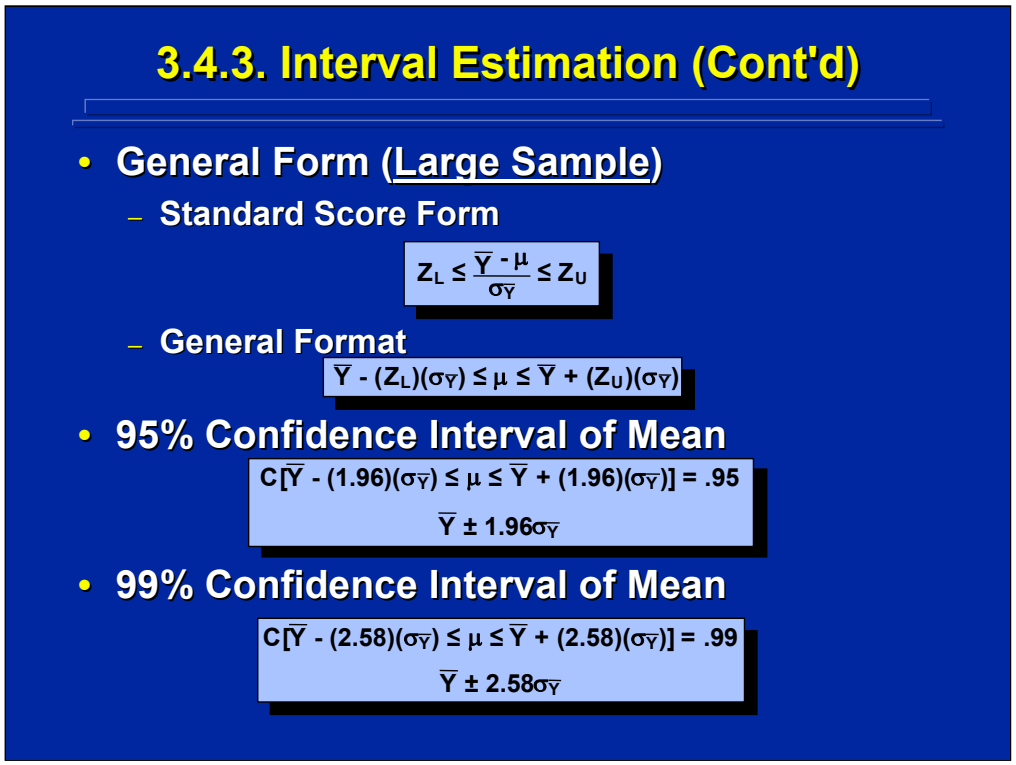

This slide summarizes some common point estimators used in experimental design. The population parameter for the mean, variance, standard deviation, and standard error (i.e., the standard deviation of the sampling distribution) are listed in the left column. The formula for the unbiased sample estimate of each population parameter is shown in the right column. Note that population parameters are stated in Greek symbols and sample point estimates are stated in Roman characters.

## 3.4.3. Interval Estimation

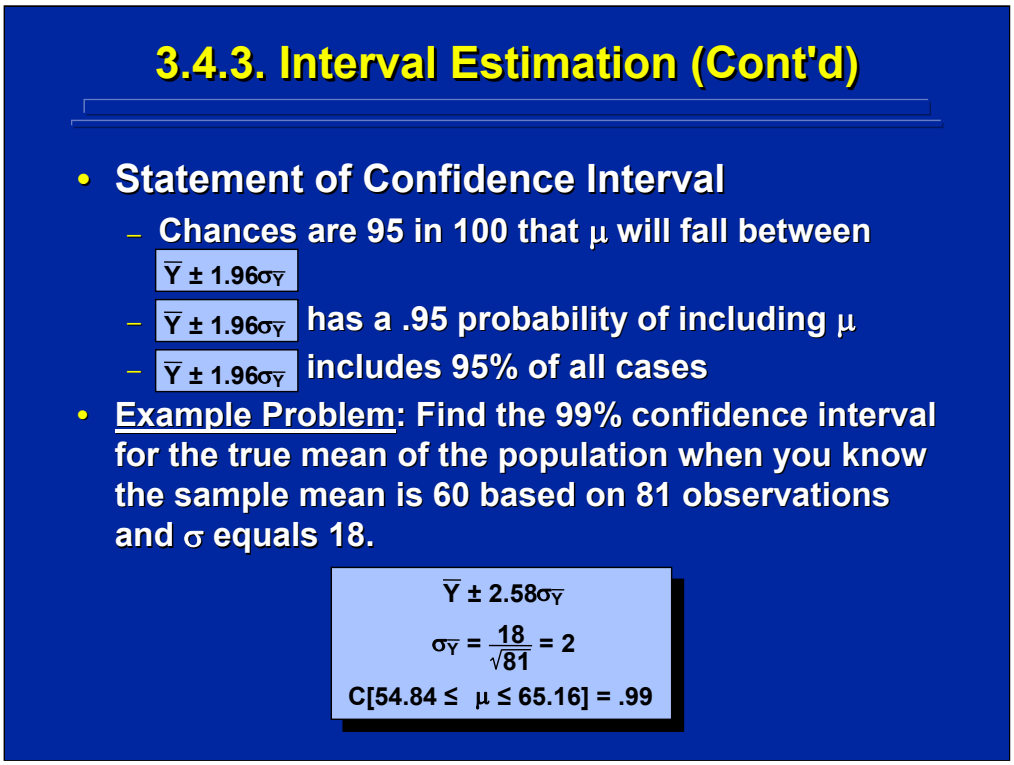
- Interval Estimation = Confidence Intervals
- Definition: A confidence interval is an estimate of a population parameter given by two numbers such that the population parameter lies between them with a certain degree of certainty.
  - Probability Statement
  - Number of Standard Deviations Above-and-Below Point Estimate
  - Based on Sampling Distribution
    - Large Sample = $\sigma$ Known = Normal Distribution
    - Small Sample = $\sigma$ Unknown = Student's t Distribution

Interval estimation is often referred to as determining confidence intervals. As defined on the slide, a confidence interval is an estimate of a population parameter given by two numbers such that the population parameter lies between them with a certain degree of accuracy. The certain degree of accuracy is the range of confidence with a lower and upper limit. First one must come up with a probability statement for the interval. An example of this is “95% confident or 99% confident.”

The number of standard deviations above and below a population parameter is used to define the interval. To calculate this interval one needs to use the appropriate sampling distribution. If one uses a large sample and the standard deviation of the population is known, then one will use the normal distribution. If one uses a small sample and the population is unknown then one will use the t distribution. A sample size of 30 is a common cut-off between large and small samples.

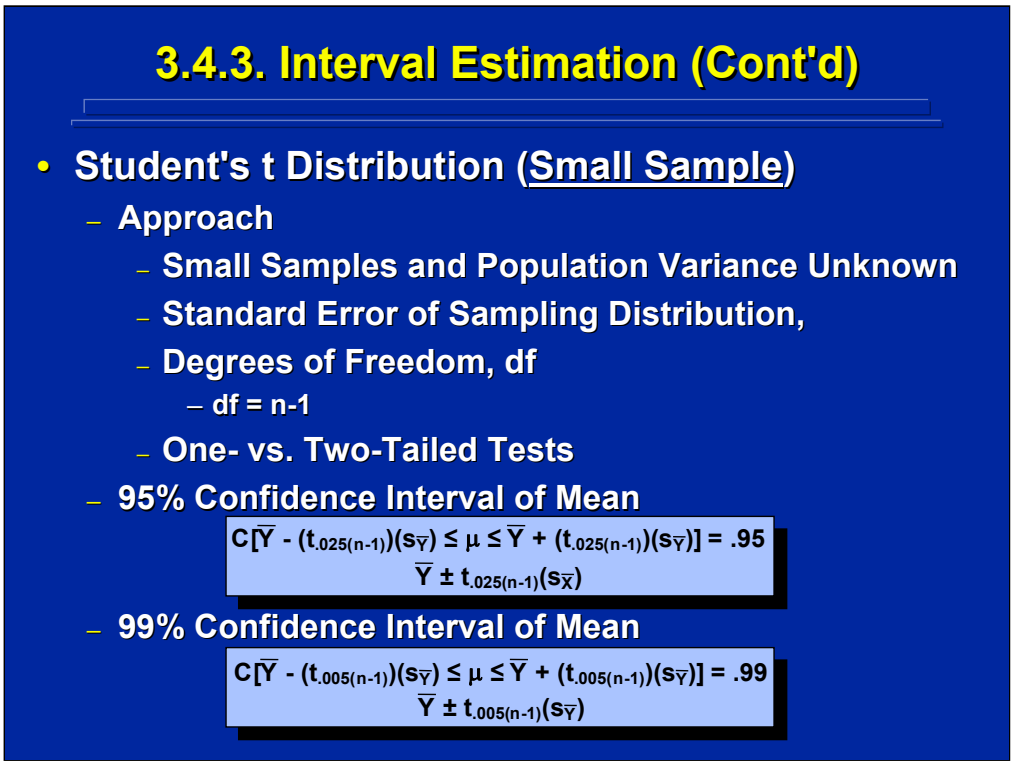



The general form of the confidence interval of the population mean, $\mu$, for a large sample is based on Z scores. In terms of stating the confidence interval of $\mu$ from the sample mean, the general format and the format for the 95% and 99% confidence intervals are shown on this slide. Notice the change in the formulas between the 95% and 99% confidence intervals is primarily the number of standard deviations of the unit normal distribution to account for 95% and 99% of the area under the curve, respectively.

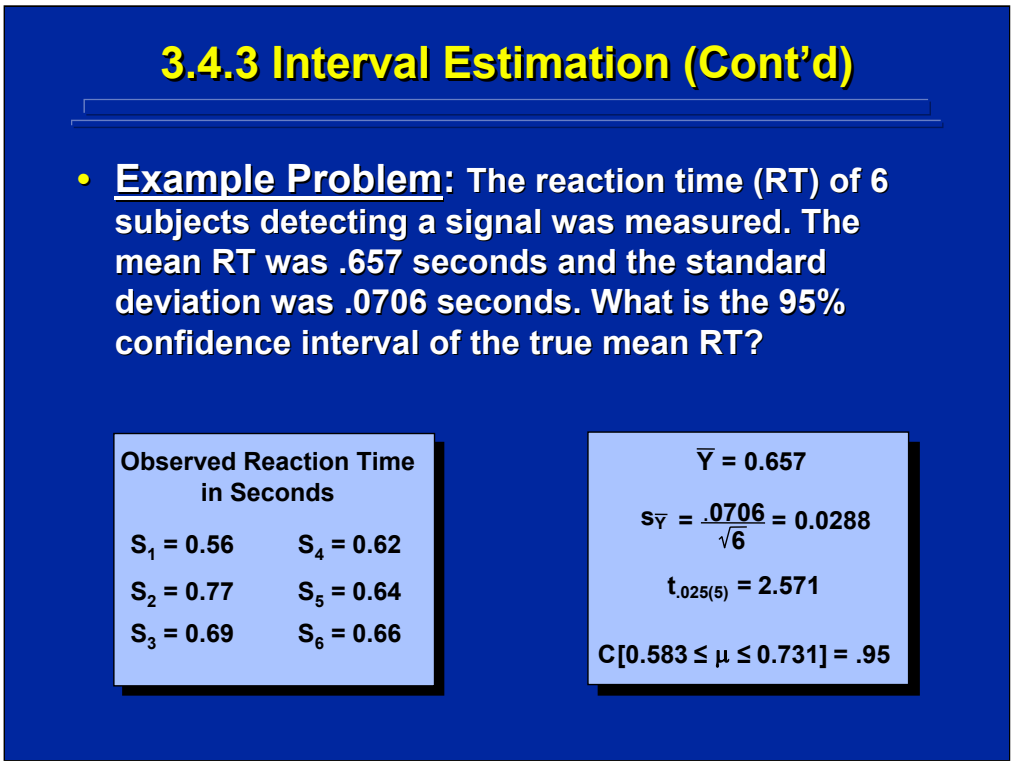



When one states a confidence interval, the experimenter makes a statement of probability. This slide shows three alternate ways of stating the same 95% confidence interval.

An example of calculating the confidence interval for the population mean is also shown on this slide. Since the sample mean is based on a large sample of 81 observations, the unit normal is the appropriate sampling distribution. The point estimate for the population mean, μ, is the sample mean, 60. The 99% confidence interval for the population mean is somewhere between 54.84 and 65.16 which is a range around the point estimate.



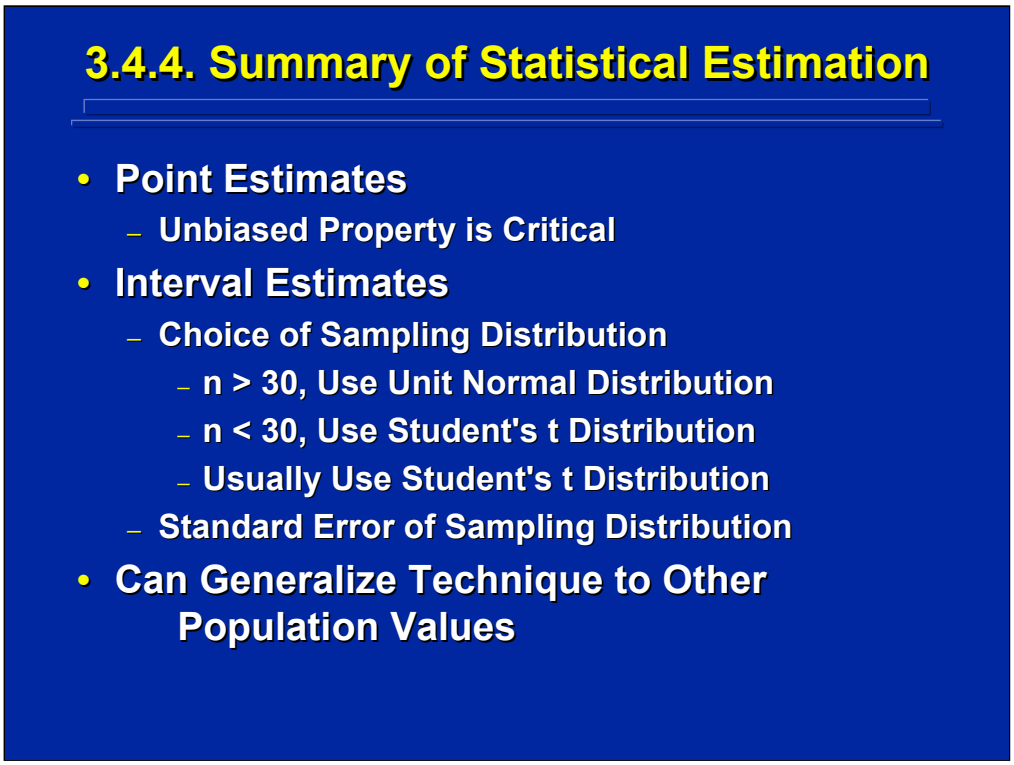

The student's t distribution is used as the sampling distribution for estimating confidence intervals based on small samples and unknown population variance. First, the standard error of the sampling distribution is calculated. To make the estimate unbiased the experimenter has to use n-1 degrees of freedom. When calculating the student's t confidence interval one should use two-tailed tabled values showing half the allowable error at the upper and lower end of the confidence interval, respectively. The general formulas for 95% and 99% confidence intervals are shown on the slide. Note that a t table is usually presented in terms of one-tailed values. Consequently, a t value of .025 is used for the 95% confidence interval, and a t value of .005 is used for the 99% confidence interval in the one-tailed, t table.




(Click in this red rectangle to see SAS calculations for this example.)

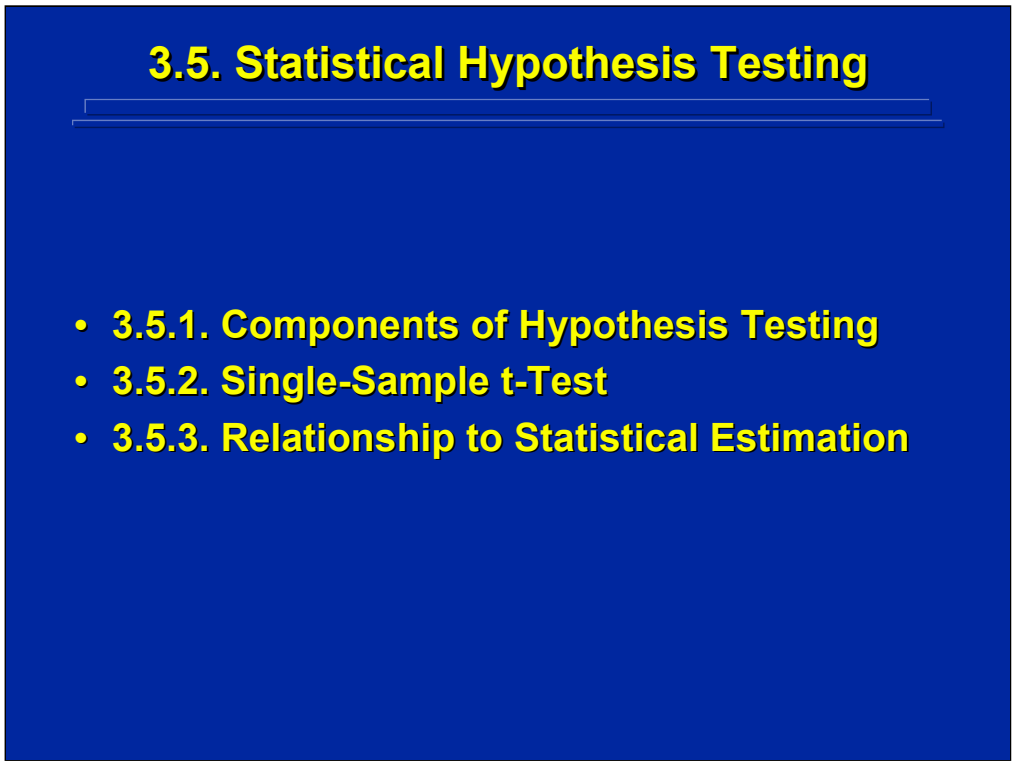
An example of calculating a small sample confidence interval of the mean based on the student's t sampling distribution is shown on this slide. The observed reaction time (RT) for each of the 6 subjects detecting a signal is shown on the left side of this slide. The mean RT was 0.657 seconds, the standard deviation was 0.0706 seconds, and the standard error was 0.0288 seconds as shown on the right side of this slide.

To find the 95% confidence interval of the true mean RT, one first finds the point estimate, then calculates the standard error, and finally finds the .025 t-value of 5 degrees of freedom from the sampling distribution. Since the t-value of 2.571 is larger than the unit normal value of 1.96, the confidence interval of the small sample estimate is larger than the large sample counterpart. Confidence intervals using the t distribution are always larger and, consequently more conservative, than the large sample confidence intervals using the Z distribution.

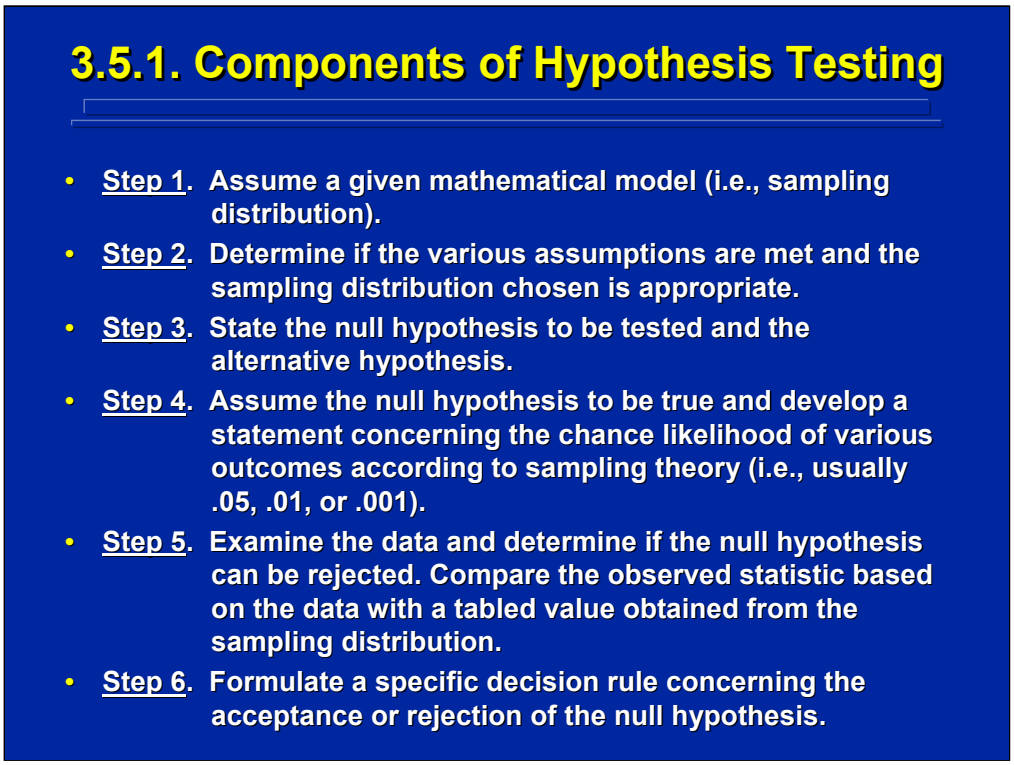



By way of summary, statistical estimation can be considered in terms of both point and interval estimation. Unbiased estimates are most critical in point estimation. Point estimates of the sample mean, variance, standard deviation, and the standard deviation of the sampling distribution (i.e., the standard error), are most often used in experimental design.

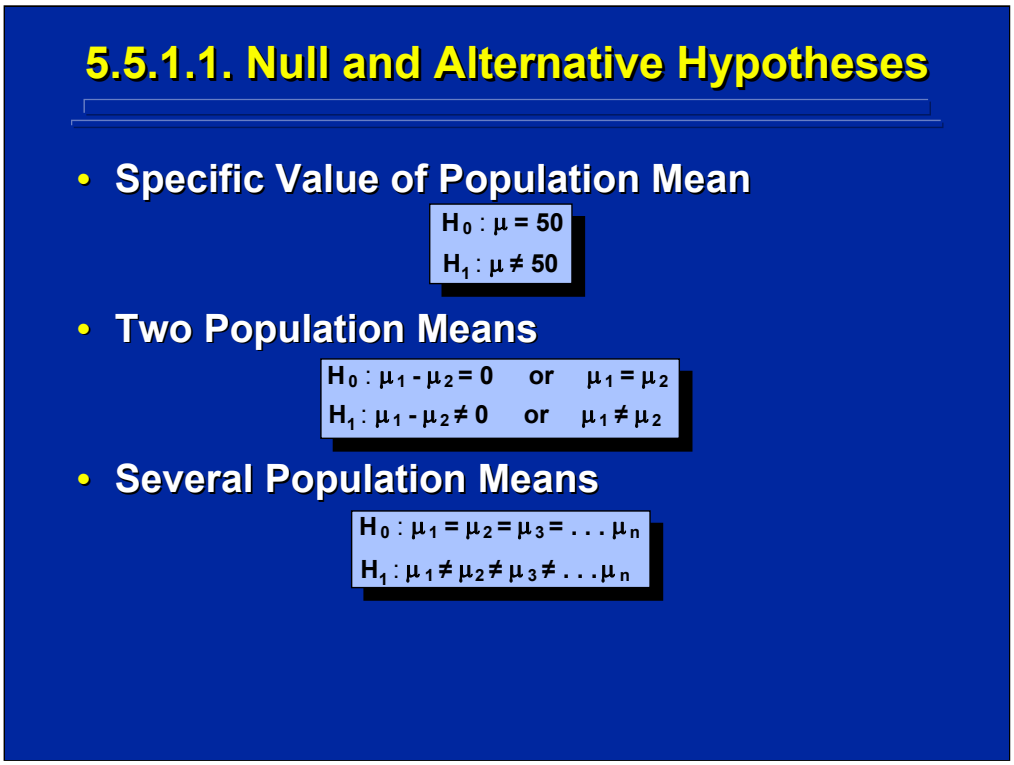
For interval estimation the choice of the appropriate sampling distribution is the key. If the sample size is greater than 30, one would use the unit normal distribution for estimating the population mean. If the sample size is less than 30, one would use the student's t distribution. In human factors research, experimenters are primarily interested in estimating the confidence interval of the population mean using the student's t distribution because small samples are typically used.



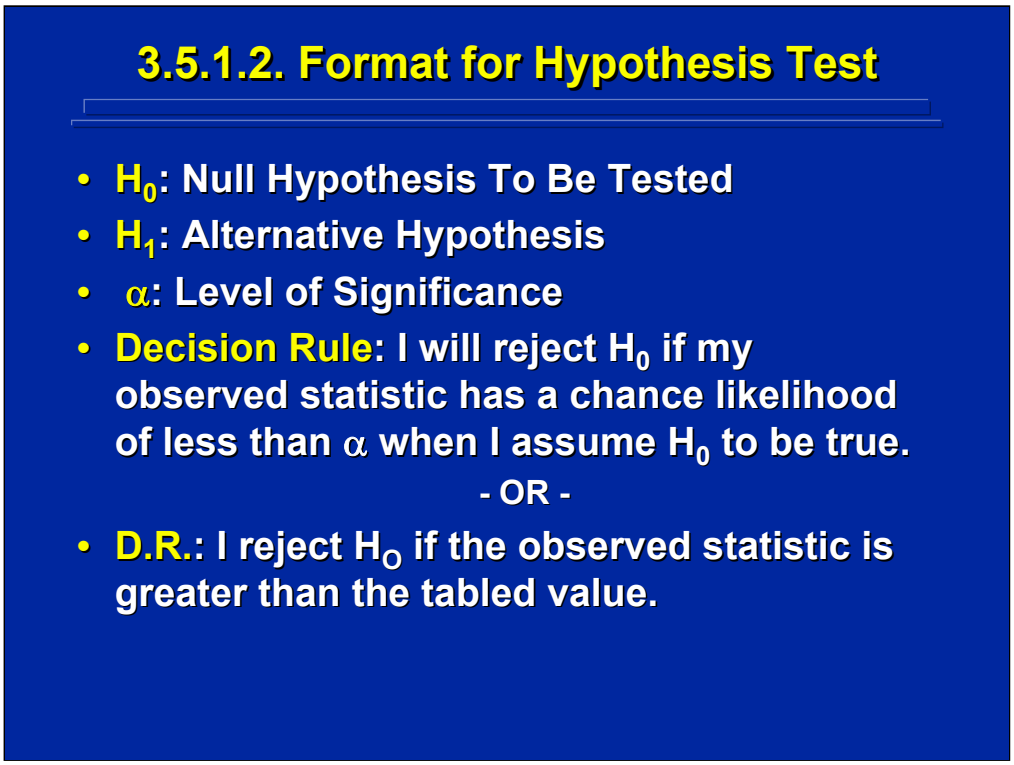

Statistical hypothesis testing is the primary inferential analysis conducted on data collected in human factors research. In this subsection, the basic components of a statistical hypothesis test are reviewed, and then these components are used in a single sample hypothesis test. In summary, the relationship between statistical hypothesis testing and statistical estimation is reviewed.



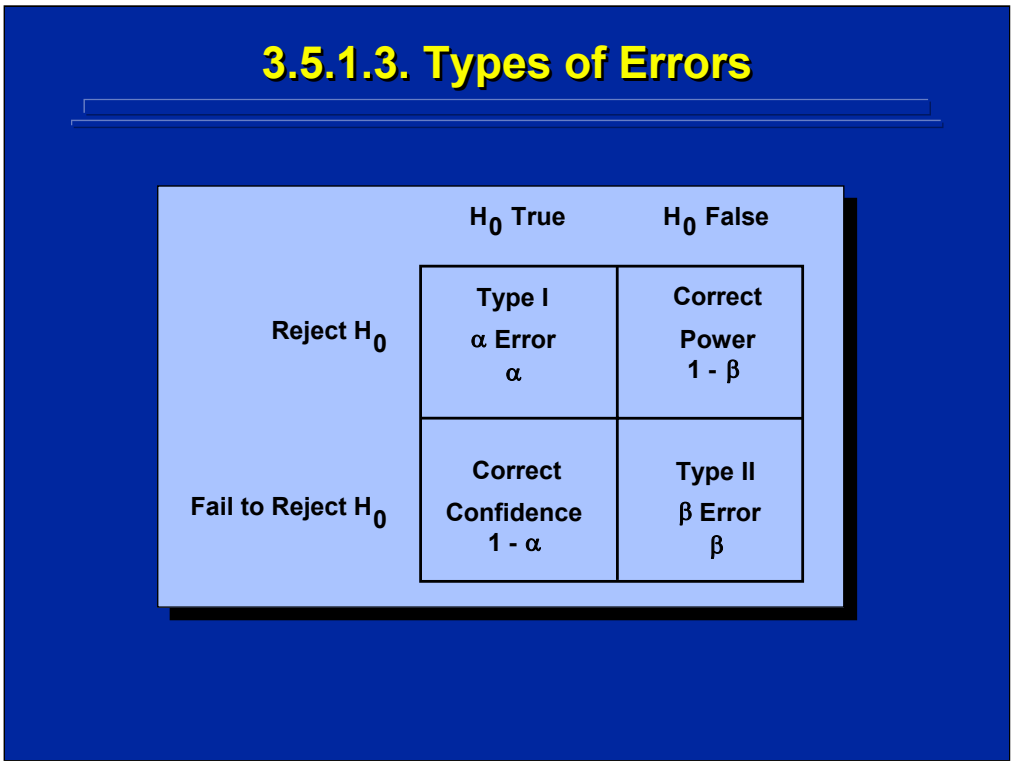

The basic components of every statistical hypothesis test can be summarized into the six steps shown on this slide. Step 1 is to select an appropriate mathematical model or sampling distribution. In human factors research experimenters primarily use the F distribution in ANOVA. Step 2 is to determine if the various assumptions of the sampling distribution are met and the sampling distribution chosen is appropriate. Step 3 is to state the null hypothesis ($H_0$), which is a statement that there is no significant difference among the treatments tested as well as the implied alternative hypothesis ($H_i$) when the null hypothesis is not true. In Step 4 one assumes the null hypothesis is true and specifies a small probability of error ($\alpha$) that one is willing to accept. The usual scientifically accepted values of alpha error are 0.05, 0.01, and 0.001. Step 5 is the comparison of the actual data collected in the experiment to the known value obtained from the sampling distribution if the null hypothesis were true. Step 6 is the formulation of a decision rule (D.R.) for accepting or rejecting the null hypothesis on the basis of the sample data collected in the research.




Depending upon the specific statistical hypothesis test that one is conducting, the null hypothesis statement of “no difference” can be expressed in terms of a specific population value, a comparison between two population means, or a comparison among several population means. This slide shows example of all three situations. Note that both the null hypothesis, $H_0$, and the alternative hypothesis, $H_1$, are always stated, explicitly or implicitly. Since experimenters primarily use two-tailed hypothesis tests in human factors research, the examples of alternative hypotheses shown on this slide are two-tailed statements of “not equal to” rather than “greater than” or “less than”.

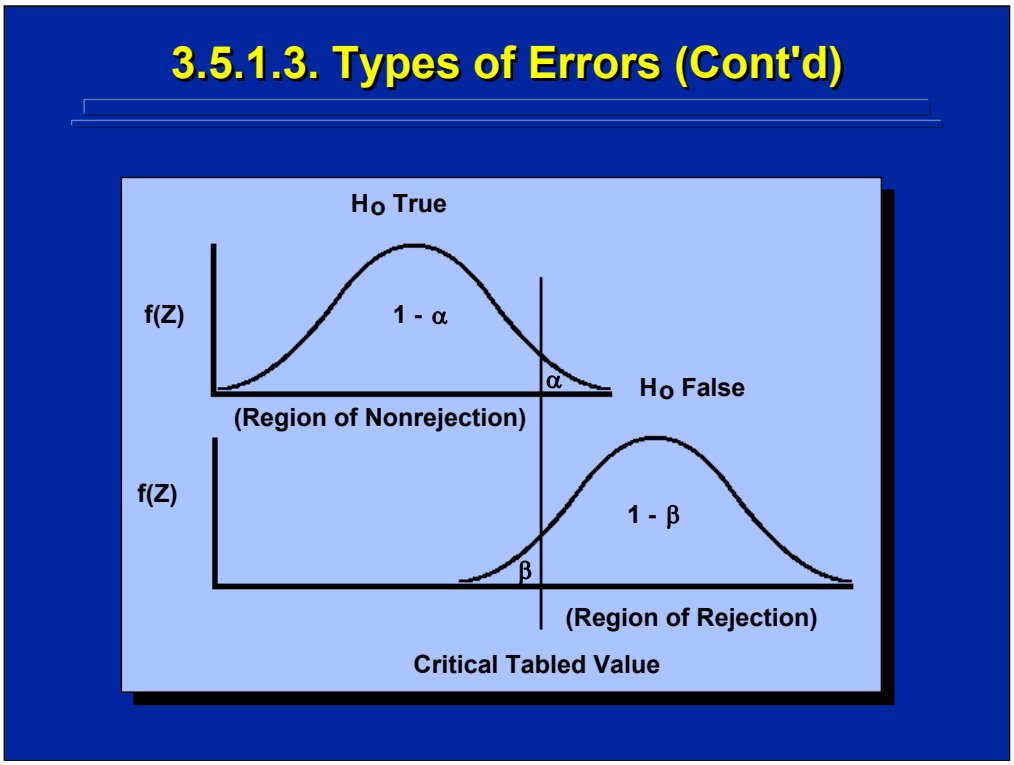



Every statistical hypothesis test can be stated in the standard format shown on this slide. This format includes four components. First one states the null hypothesis, $H_0$, to be tested. Second, one states the alternative hypothesis, $H_1$, usually as a two-tailed test. Third, one states the level of significance, $\alpha$, set by the experimenter. And, fourth one states the decision rule (D.R.) for rejecting the null hypothesis. The simplest way of stating the D.R. is saying that the null will be rejected if the observed statistic calculated from the data collected in the experiment is greater than the tabled value drawn from the appropriate sampling distribution. Although this format is implicit and usually not stated in every statistical hypothesis test, it will be used throughout this reference for emphasis.




In the true state of nature the null hypothesis is either true or false. The resulting 2x2 contingency table of possible decision outcomes is shown on this slide. There are two ways of making errors and two ways of being correct. If the null hypothesis is true but one rejects it, a Type I or alpha error occurs with a probability of $\alpha$. The experimenter directly sets $\alpha$ error at some small level such as 0.05, 0.01, or 0.001 before conducting the hypothesis test. If the null hypothesis is false and the experimenter fails to reject it, then a Type II or beta error occurs with a probability of $\beta$.

One would be correct if the null hypothesis is true and the experimenter fails to reject it. This is called the level of confidence of a test and has a probability of $1-\alpha$. In statistical hypothesis testing, one strives to have a high level of confidence in the test by setting $\alpha$ error (i.e., level of significance) low. One would be correct if the null hypothesis is false and it is rejected. This is called the power of a test, and it has a probability of $1-\beta$. One strives to conduct the most powerful test as possible while maintaining a high level of confidence that a true difference exists.




This is a graphical representation using the unit normal sampling distribution of the 2x2 contingency table shown in the previous slide. The upper probability distribution occurs when the null hypothesis is true, and the lower distribution represents a situation when the null hypothesis is false. The experimenter chooses the level of significance of the statistical test by choosing a level of alpha error. This is depicted by moving the vertical line shown in the figure to the left or right.

One wants to have the largest degree of confidence and the greatest statistical power in a hypothesis test. But, this requires a tradeoff. In order to have a high level of confidence (1-$\alpha$), the experimenter must choose a low probability of $\alpha$ error. Note that $\alpha$ error is set directly by the experimenter, but it indirectly changes $\beta$ error and power (1-$\beta$) of the statistical test. If one reduces the confidence then one increases the power, and vice versa.

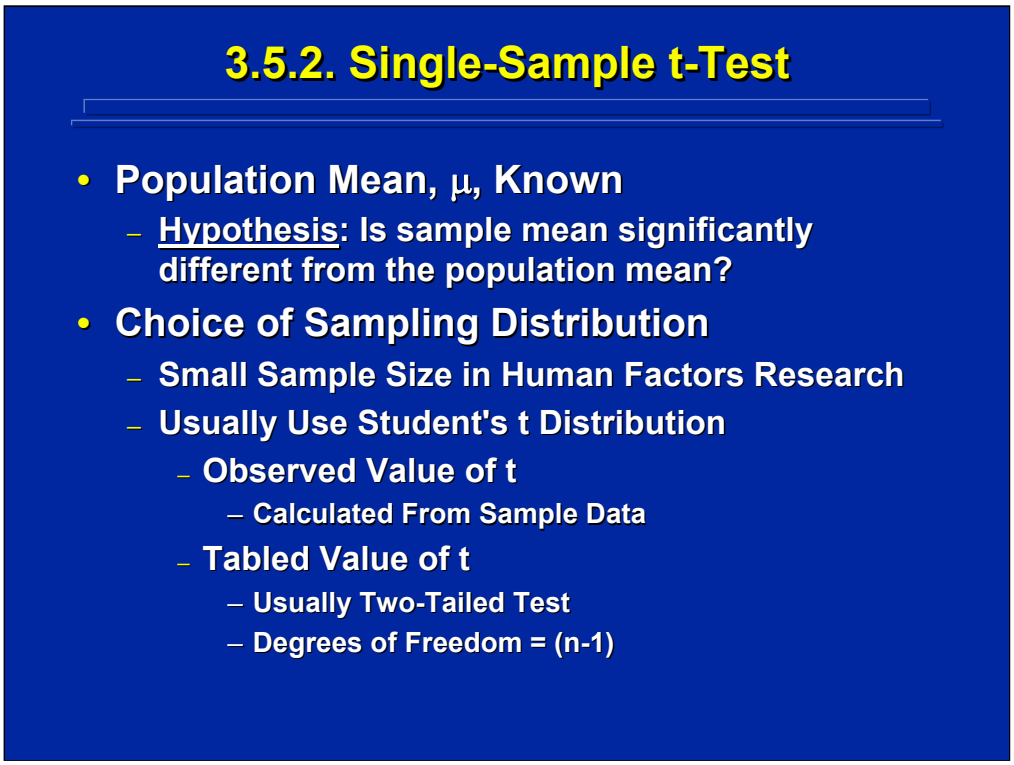
During statistical hypothesis testing, the experimenter makes a simple decision based on the evidence in the sample data to either reject or fail to reject the null hypothesis. Rather than accept the null hypothesis, one should fail to reject it. This is done in order to avoid a Type II error shown on this slide.

## 3.5.1.4. Statistical Power

- **Definition: Probability of rejecting the null hypothesis when the null hypothesis is false.**
- **Experimenter Sets $\alpha$ Directly**
- **Power Indirectly Affected by**
  - **$\alpha$ Level**
  - **Population Variance**
  - **Sample Size**
- **Statistical Power Calculations**
  - **$\alpha$ Level and Estimate of Variance Known**
  - **Solve for Sample Size Required**

Statistical power is the probability of correctly rejecting the null hypothesis when it is false and is affected by $\alpha$ error, the population variance, and the sample size. The researcher can influence the statistical power of a hypothesis test when choosing both $\alpha$ and sample size. As the experimenter chooses a smaller $\alpha$ error, the power (1-$\beta$) of the hypothesis test also decreases since $\beta$ error indirectly increases. By changing sample size, n, the experimenter can indirectly change the power. In general, the larger the sample size, the more powerful the hypothesis test becomes.

Statistical power calculations can help the experimenter determine the sample size for the experiment. If one has an estimate of the population variance and has made a decision on the $\alpha$ value of the hypothesis test, the sample size needed for a given level of power can be determined. If cost and time are not major constraints in choosing sample size, one could use a power analysis to determine the appropriate sample size.




A single sample hypothesis test is used to demonstrate the logic of a statistical hypothesis test. In this example, the experimenter is conducting a hypothesis test to determine if a sample mean is significantly different from a known population value, μ.

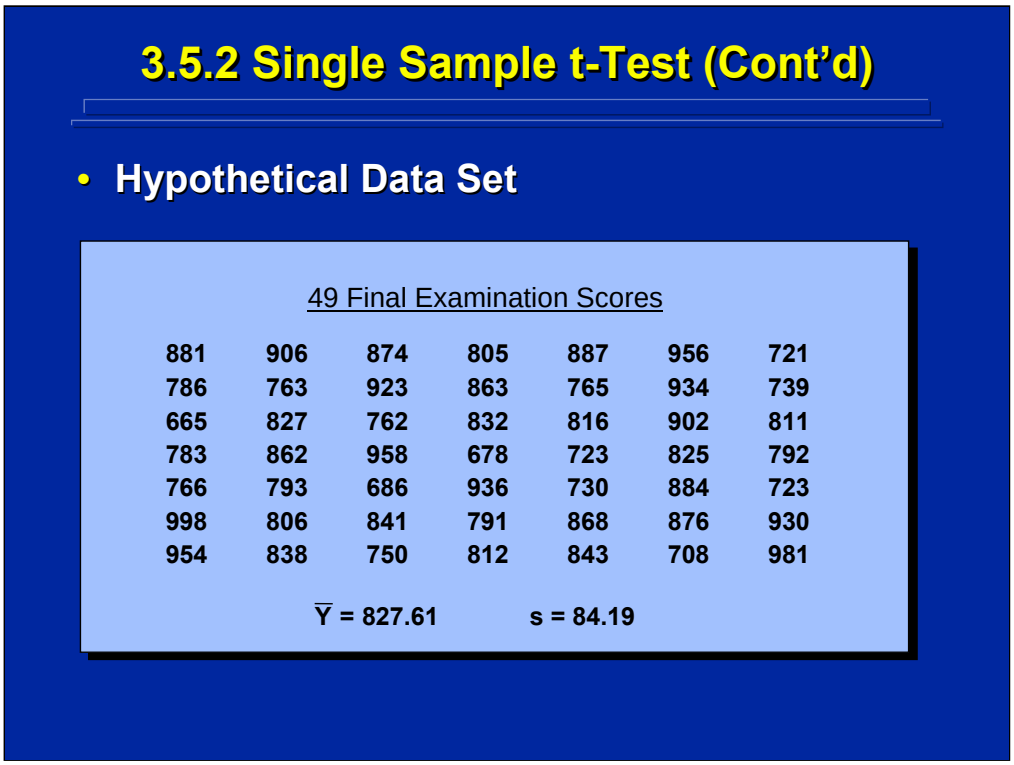
Choice of the appropriate sampling distribution is fundamental to any statistical hypothesis test. Since human factors researchers primarily use small samples, they usually use the t rather than the Z distribution when they have a choice between them. The actual statistical test consists simply of calculating the t-observed value based on the sample data and comparing it to the t-tabled value drawn from the t distribution with the appropriate degrees of freedom.

## 3.5.2. Single-Sample t-Test (Cont'd)

- **Example Problem: The experimenter wishes to compare the average scores on the final examination in a military course to a standard population value of 792 points for course mastery. Forty-nine trainees are randomly assigned to a particular section of the course. The experimenter is interested in determining if the average score on their final examination is significantly different from the known mastery value of 792 points at the .05 level of significance.**

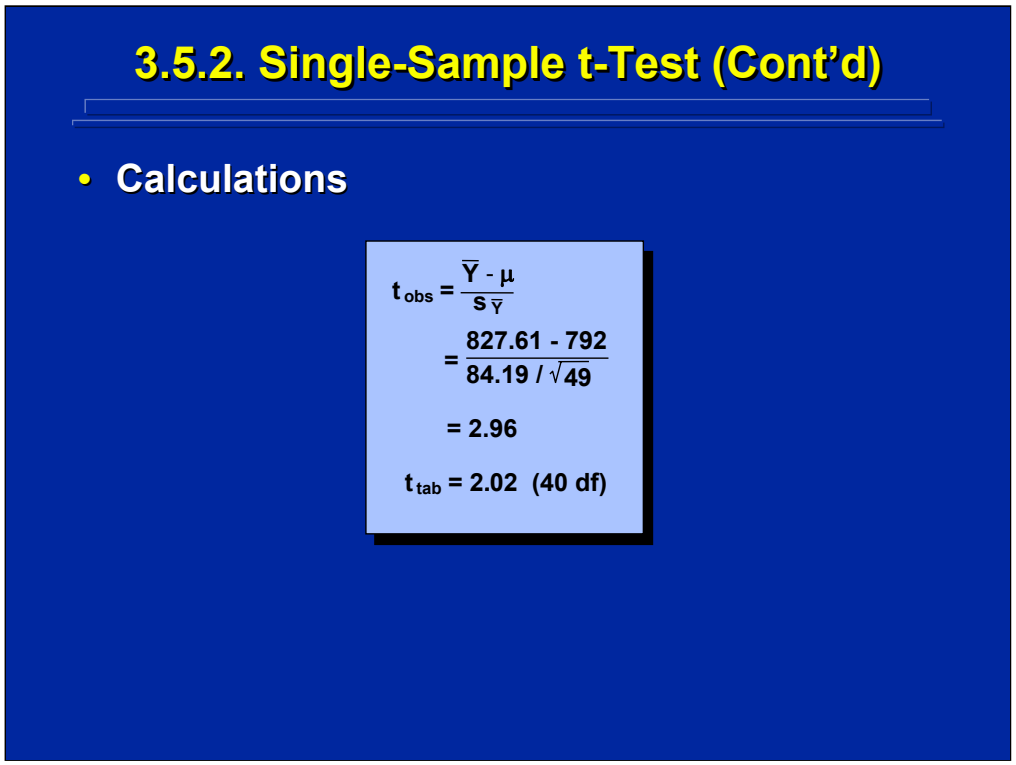
(Click in this red rectangle to see SAS calculations for this example.)

An example problem of a single sample hypothesis test is given on this slide. One is trying to determine if the final examination of the sample is significantly different from 792 points that represents the known average for students who have mastered the course material. Notice this example is stated as a two-tailed test, because it is testing only that the sample is significantly different from a known value of 792 and not testing the direction of difference.



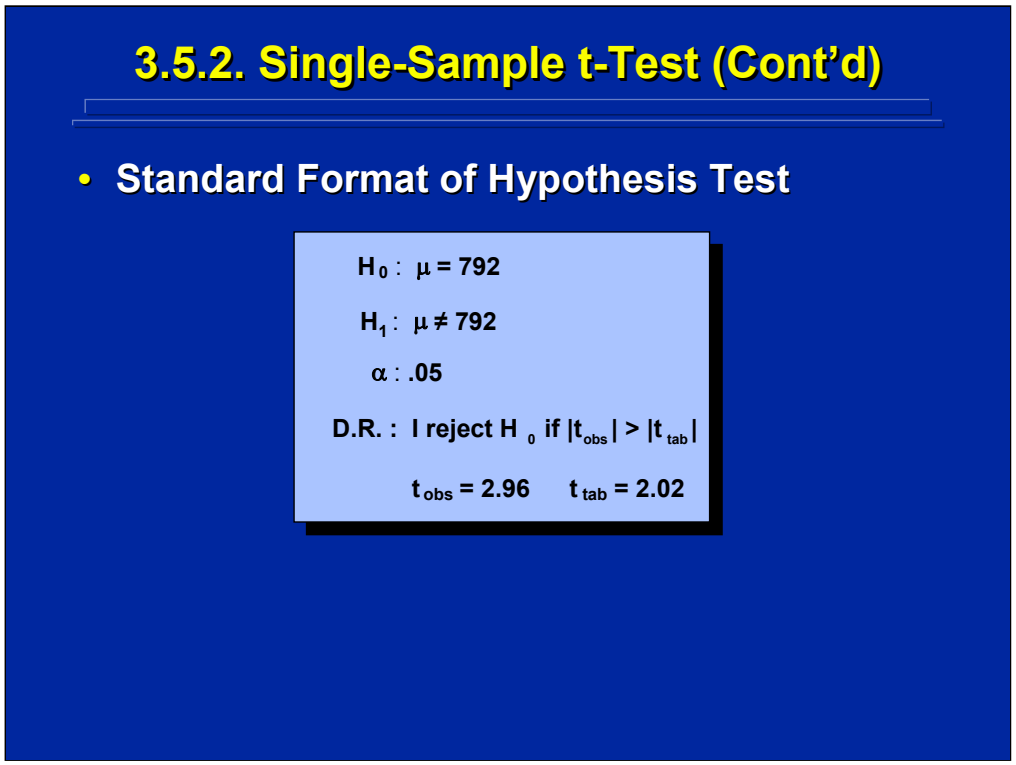

(Click in this red rectangle to see SAS calculations for this example.)

This is a listing of the 49 final examination scores obtained from the class. The average score on the final exam is 827.61 and the standard deviation is 84.19.



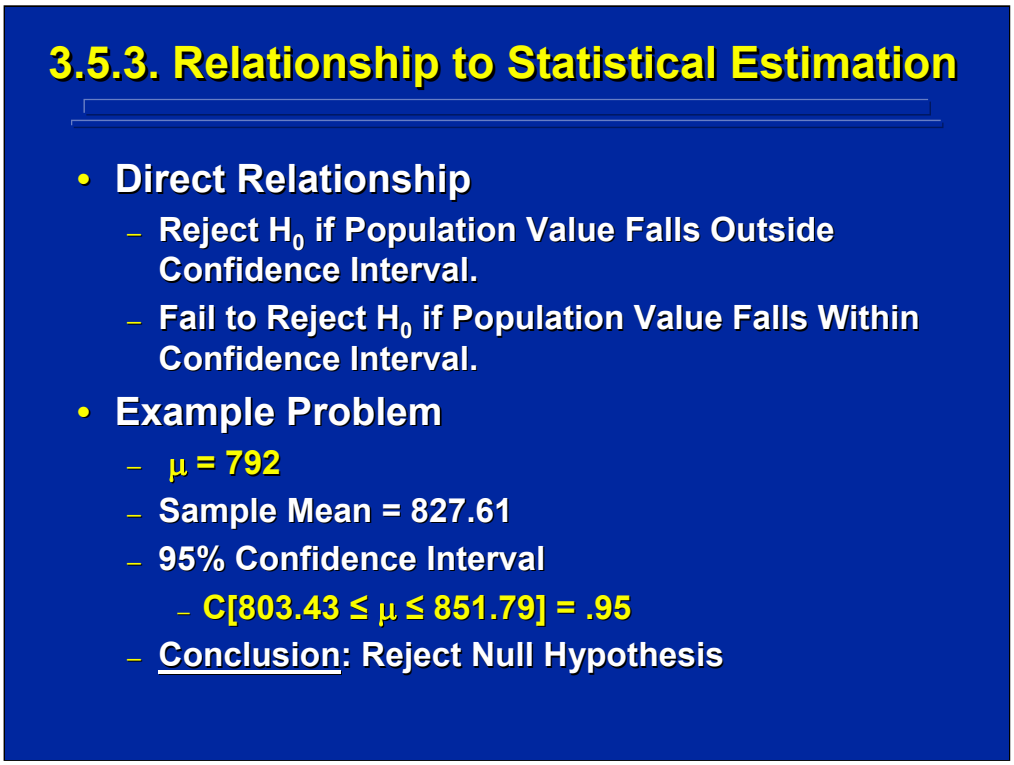

(Click in this red rectangle to see SAS calculations for this example.)

Although the sample size, 49, is large enough to use the unit normal sampling distribution, one could be conservative and use the t-distribution. The actual calculations of t-observed and t-tabled for the example problem are shown on this slide. When using standard tables to find the t value, only 40 and 50 degrees of freedom are listed, not 48. Since the t distribution is not linear, one cannot use a linear interpolation of tabled values. The conservative approach is to use the lower of the two values around the actual degrees of freedom. In this case, the tabled value of 40 degrees of freedom would be used instead of 50. This makes it more difficult to reject the null hypothesis and is, therefore, more conservative.



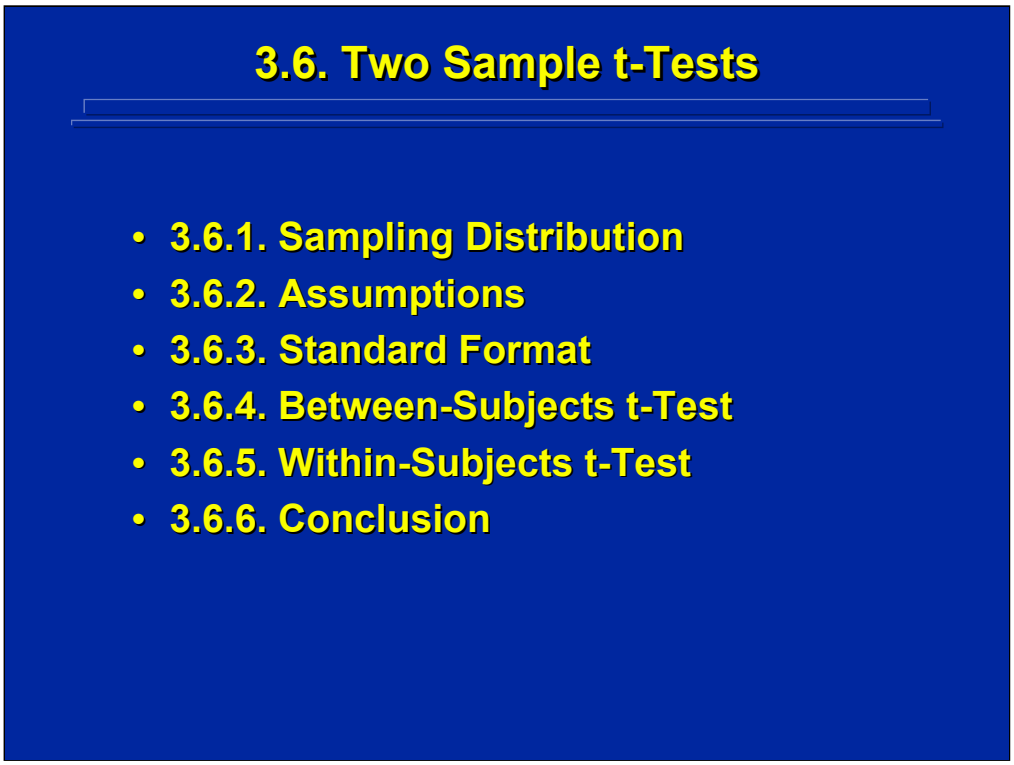

(Click in this red rectangle to see SAS calculations for this example.)

The hypothesis test can be summarized in standard format. Note that the alternative hypothesis is stated as a two-tailed test. Likewise, the decision rule uses absolute values for a two-tailed test (p = .025), and the observed and tabled values are stated as t-values from the t sampling distribution. The final decision for this example is to reject the null hypothesis since the observed value is greater than the tabled value.

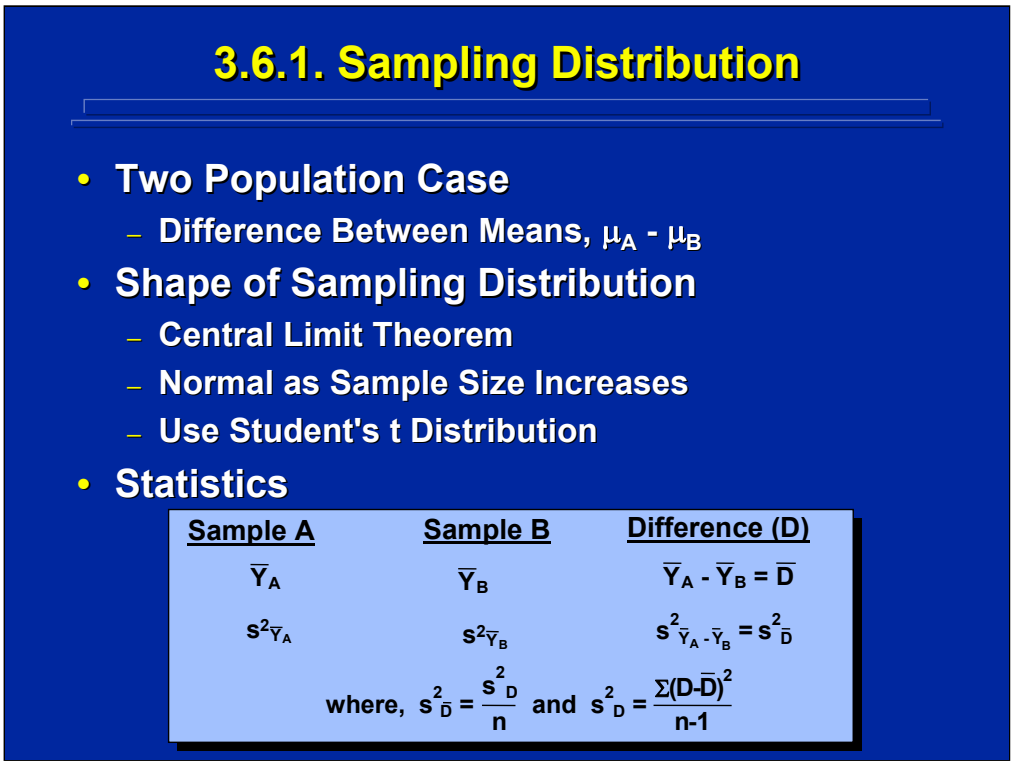



(Click in this red rectangle to see SAS calculations for this example.)

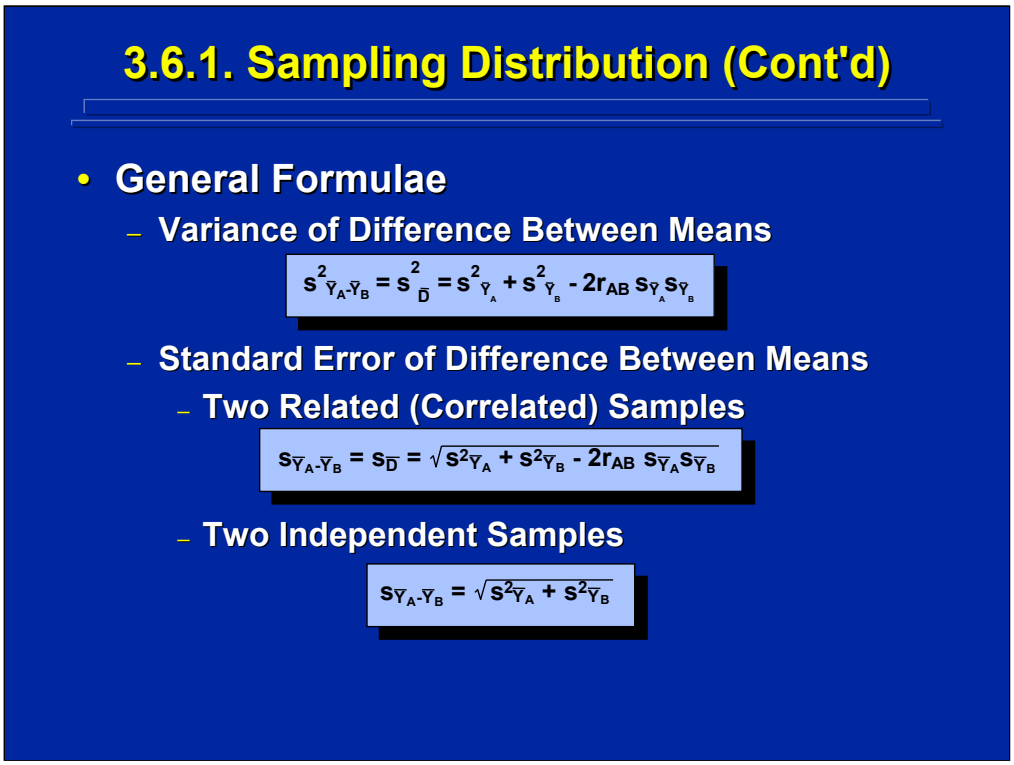
There is a direct relationship between statistical hypothesis testing and estimation. If the population value falls outside of the confidence interval estimated by the sample, then one rejects the null hypothesis. Otherwise, one fails to reject the null hypothesis. In this example, the population value of 792 falls outside the 95% confidence interval, and the null hypothesis was rejected.




A test of significant difference between two sample means is probably the most common type of statistical hypothesis test. As in any hypothesis testing, the experimenter must choose the appropriate sampling distribution, consider the assumptions, and develop the standard testing format. A t-test is commonly used to compare two sample means, and the procedure for calculating the observed t-value differs depending upon whether the two samples are independent or related. Basic concepts of two sample t-tests are reviewed, and computational examples of between-subjects and within-subjects t-tests are presented in this subsection.




The sampling distribution for two-sample hypothesis tests is a sampling distribution based on the difference between two means. Based on the central limit theorem, this sampling distribution will be normally distributed for large sample sizes. However, since human factors researchers primarily use small sample sizes, they usually use the t distribution instead of the normal distribution for tests of differences between two means. In addition to the mean and standard error of Sample A and B, the formula for the difference scores, D, between the two samples and the standard error of D are shown in the bottom portion of this slide. Hypothesis tests of differences between the means of two samples use all of these statistics.




When testing the significant difference between two means, one must calculate the variances of difference as well as the standard error of differences. The standard error is the standard deviation of the sampling distribution of differences between means. As shown on this slide, the standard error of differences is calculated differently depending on the relationship between the two samples. If the samples are related (i.e., a within-subjects design), then correlation between samples, $r_{AB}$, is included in the formula. If the samples are independent (i.e., a between-subjects design), then the correlation aspects of the formula is 0 and does not appear.

## 3.6.2. Assumptions

- Normality
  - Samples Drawn from Normal Distributions
    - Robust to Violation
      - Equal Sample Size
      - Two-Tailed Test
- Homogeneity of Variance, $\sigma^2_A = \sigma^2_B$
  - Ignore with Equal Sample Size
  - Preliminary F-Test when $n_A \neq n_B$
- Sample Relationship
  - Independent Samples
  - Related Samples
- Population Means are Equal, $\mu_A = \mu_B$
  - Null Hypothesis

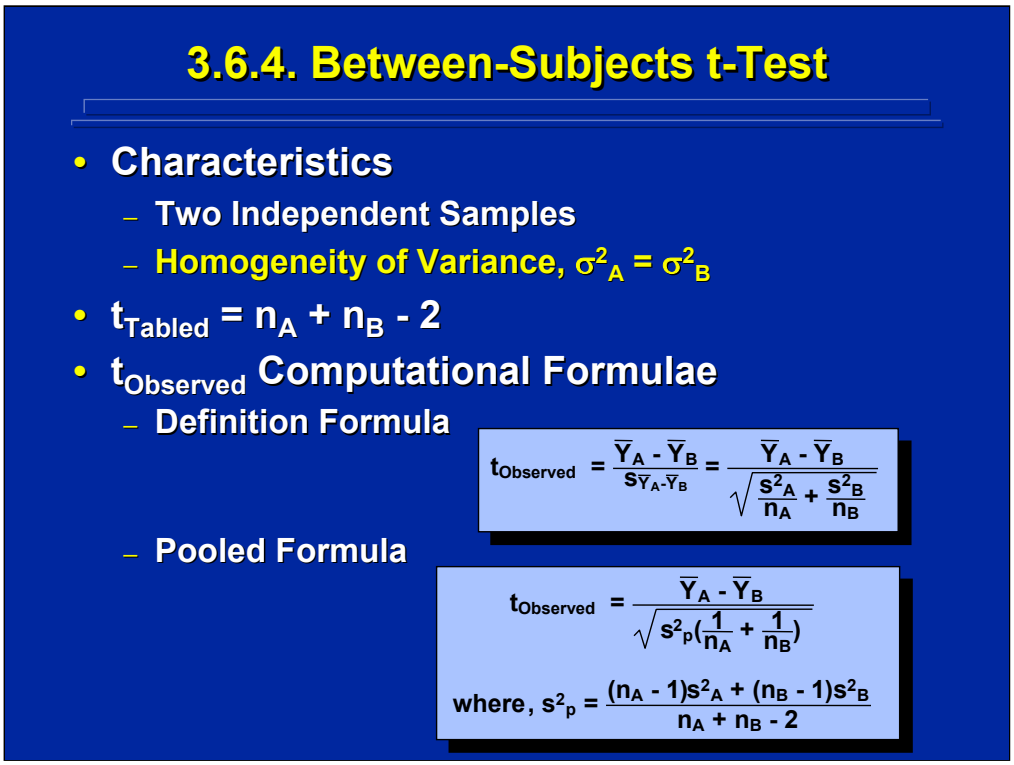
There are four critical assumptions that must be considered in two sample t-tests. The first is that both samples are drawn from normal distributions. The t-test is robust to a violation of this assumption if one uses equal sample sizes and a two-tailed test. Robustness means that the sampling distribution can be used to determine the tabled value even though the assumption is not met. A second assumption is homogeneity of variance. This means that the population variance of one sample mean is equal to the population variance of the other sample. Again, a t-test is robust to violation of this assumption if the sample sizes are equal. If sample sizes are not equal, then one should use a preliminary F-test to test the homogeneity of variance assumption. The third assumption deals with the relationship of the two samples. If a different random sample of subjects is used for each sample (i.e., between-subjects samples), then the samples can be assumed to be independent. If the same subjects (i.e. within-subjects samples) or matched subjects are used, then the two samples are correlated. Formulae for t-observed vary for between-subjects and within-subjects t-tests. The fourth assumption is that the population means are equal. This last assumption is really the null hypothesis that is being tested in the t-test itself.

## 3.6.3. Standard Format

- **Test Format**
  - $H_0$: $\mu_A = \mu_B$
  - $H_1$: $\mu_A \neq \mu_B$
  - $\alpha$: .05, .01, or .001
  - D.R.: I reject $H_0$ if $|t_{Observed}| > |t_{Tabled}|$
    - $t_{Tabled\ (\alpha/2)} = (n_A + n_B - 2)\ df$
    - $t_{Observed} = (\bar{Y}_A - \bar{Y}_B) / s_{\bar{Y}_A - \bar{Y}_B}$
- **Alternatives for the $t_{Observed}$ Statistic**
  - Between-Subjects Samples
    - Homogeneity of Variance
    - Heterogeneity of Variance
  - Within-Subjects Samples

The standard format for stating the null and alternative hypotheses, level of significance, and the decision rule for a two sample t-test is shown at the top of portion of this slide. Note that the four standard components of the test format are listed, and the specifics of each component are tailored to the particular hypothesis test being conducted.

For a two sample t-test, the null hypothesis states that the two population means are equal. The alternative hypothesis is a two-tailed test stating that the two means are “not equal” but not the direction of the difference. In order to maintain robustness to violations of assumptions, one usually uses two-tailed tests. The particular $\alpha$ level is selected by the experimenter and generally depends on the consequences of falsely rejecting a true null hypothesis. The decision rule is to reject the null hypothesis if the t-observed value is greater than the t-tabled value. Since each of the two samples loses 1 degree of freedom, the t-tabled value is based on losing 2 degrees of freedom. Note that the t-tabled value equals $\alpha/2$ in a one-tailed t-table. The specific formula for calculating the t-observed value differs for between-subjects and within-subjects samples. Each is described separately.




The formulae on this slide represent the calculations for t-observed for a between-subjects t-test in which one has equal sample size and two-tailed tests. In this situation, homogeneity is usually not tested due to the robustness of the t-test. Both the definition and pooled formulae are shown on the slide. They are algebraically equivalent. So, either formula can be used to calculate the t-observed statistic from the data collected in the experiment.

## 3.6.4. Between-Subjects t-Test (Cont'd)

- **Characteristics**
  - Two Independent Samples
  - Heterogeneity of Variance, $\sigma^2_A \neq \sigma^2_B$
  - Preliminary F-Test when $n_A \neq n_B$
- **Hartley $F_{max}$ Test for Homogeneity of Variance**
  - $H_0$: $\sigma^2_A = \sigma^2_B$
  - $H_1$: $\sigma^2_A \neq \sigma^2_B$
  - $\alpha$: .20
  - D.R.: I reject $H_0$ if $F_{Observed} > F_{Tabled}$
    - $F_{Observed} = s^2_{Larger} / s^2_{Smaller}$
    - $F_{Tabled}$ = Hartley $F_{max}$ in Winer et al. (1991) Table D.7 where n = $n_{Largest}$ and k = 2

When sample size is not equal, one often conducts a preliminary test for heterogeneity of variance. The $F_{max}$ test can be used in this situation (Winer, Brown & Michels, 1991, pp. 104-105). This slide shows the standard format for a preliminary test of the assumption of equal population variance. Note that this test is trying to accept the null hypothesis of no difference (i.e., homogeneity of variance). Consequently, one chooses a high $\alpha$ error (i.e. 0.20) to guard against Type II, or $\beta$, error indirectly.

The F-observed value is determined by the ratio of the larger sample variance divided by the smaller sample variance. The $F_{max}$ tabled value is found in Table D.7 of Winer et al. (1991) by using the degrees of freedom of the largest sample size and k =2 for the comparison of the 2 variances used in the t-test. If there is a significant difference, then heterogeneity of variance exists, and the homogeneity of variance assumption is violated.

## 3.6.4. Between-Subjects t-Test (Cont'd)

- **Heterogeneity of Variance**
  - $t_{Observed}$ Computational Formula

$$t_{Observed} = \frac{\bar{Y}_A - \bar{Y}_B}{s_{\bar{Y}_A - \bar{Y}_B}} = \frac{\bar{Y}_A - \bar{Y}_B}{\sqrt{\frac{s^2_A}{n_A} + \frac{s^2_B}{n_B}}}$$

  - $t_{Tabled}$ = Cochran and Cox (1957) $t'_{Adjusted}$ Formula

$$t'_{Adjusted} = \frac{s^2_{\bar{Y}_A} t_A + s^2_{\bar{Y}_B} t_B}{s^2_{\bar{Y}_A} + s^2_{\bar{Y}_B}}$$

$t_A = t_{Tabled}$ with $(n_A - 1)$ df

$t_B = t_{Tabled}$ with $(n_B - 1)$ df

When heterogeneity exists the normal between-subjects t-test cannot be used because the observed t calculated from the two samples is not truly distributed according to the t distribution. One calculates the t-observed value as usual. But, Winer, et al. (1991, pp. 67-69) recommend that one can use the Cochran and Cox (1957, p. 101) t' adjustment for the t-tabled value by using the formula shown on this slide.

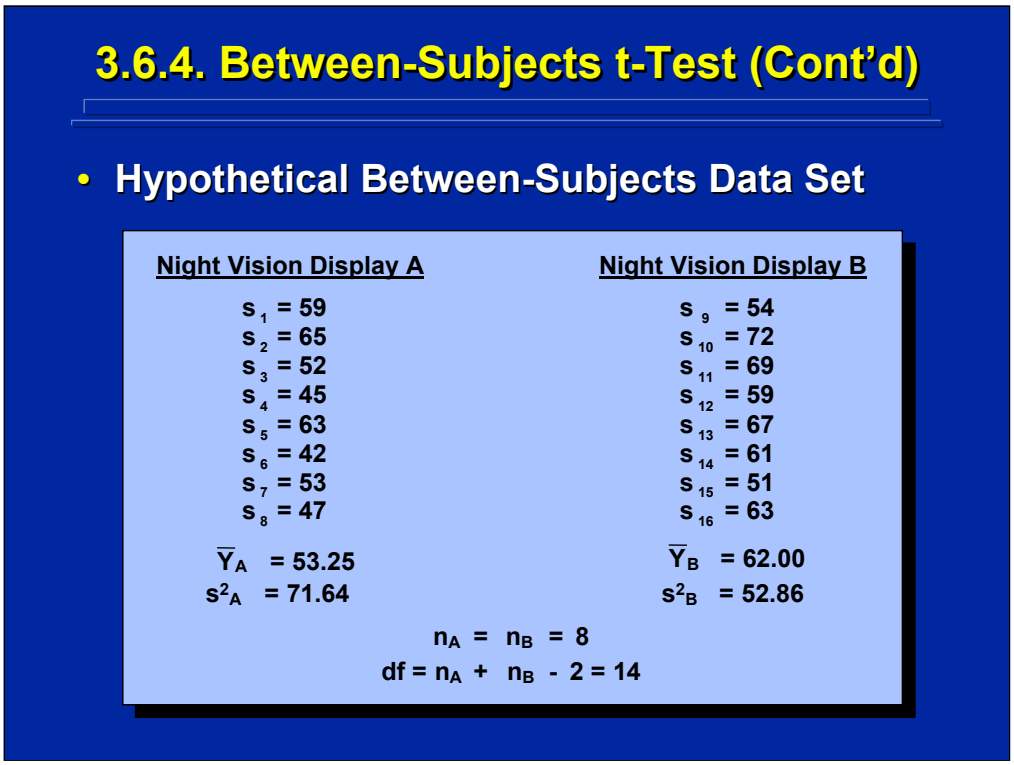
## 3.6.4. Between-Subjects t-Test (Cont'd)

- **Example Problem: An experimenter wishes to compare performance of two different night vision displays used in nighttime maneuvering. Eight squads used display A, and eight different squads used Display B. Each squad completed the same nighttime maneuver. The experimenter wants to determine if there is a significant difference ($p < 0.05$) in mean time in minutes to complete the nighttime maneuver between using the two night vision displays.**

(Click in this red rectangle to see SAS calculations for this example.)

This is a between-subjects design because 8 squads used night vision Display A, and 8 different squads used night vision Display B while performing the nighttime maneuver. Since the sample size is only 8, the t-distribution is the appropriate sampling distribution for testing the difference between minutes to complete the nighttime maneuver while using the two night vision displays.

A two-tailed t-test is used since the experimenter is interested in any significant difference between the two night vision displays regardless of direction. The standard t table presents only one-tailed values. Consequently, the t-tabled value is set at $\alpha = 0.25$ to yield a two-tailed test at the at $p < 0.05$.




(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the basic layout of the two-group experiment that compares night vision displays A and B. Note that the eight different squads in each of the two treatment conditions are designated by a different subscript number yielding a total of 16 different squads used in this two group experiment. Sample size refers to the number of squads using each night vision display, not the total number of different squads. In this example an equal sample size (n) of 8 squads is used. One degree of freedom is lost in each of the two display conditions resulting in a t-test with 14 df as shown on the slide.

Since different subjects were used in each display condition and sample size is equal, this hypothesis test is a between-subjects t-test in which homogeneity of variance usually does not need to be tested beforehand. Note that the SAS analysis as shown in Slater and Williges (2006) appendix automatically conducts the preliminary test for homogeneity of variance. The SAS test shows that the difference between the sample variances (i.e., the square of the two sample standard deviations shown in the SAS output) of 71.64 and 52.86 is not significant at the 80% level of significance (i.e., $\alpha$ = 0.20). Hence, one can assume homogeneity of variance and a subsequent valid t-test between means.

## 3.6.4. Between-Subjects t-Test (Cont'd)

- Calculation of $t_{Observed}$ Using Pooled Formulae

$$s^2_p = \frac{(n_A - 1)s^2_A + (n_B - 1)s^2_B}{n_A + n_B - 2}$$

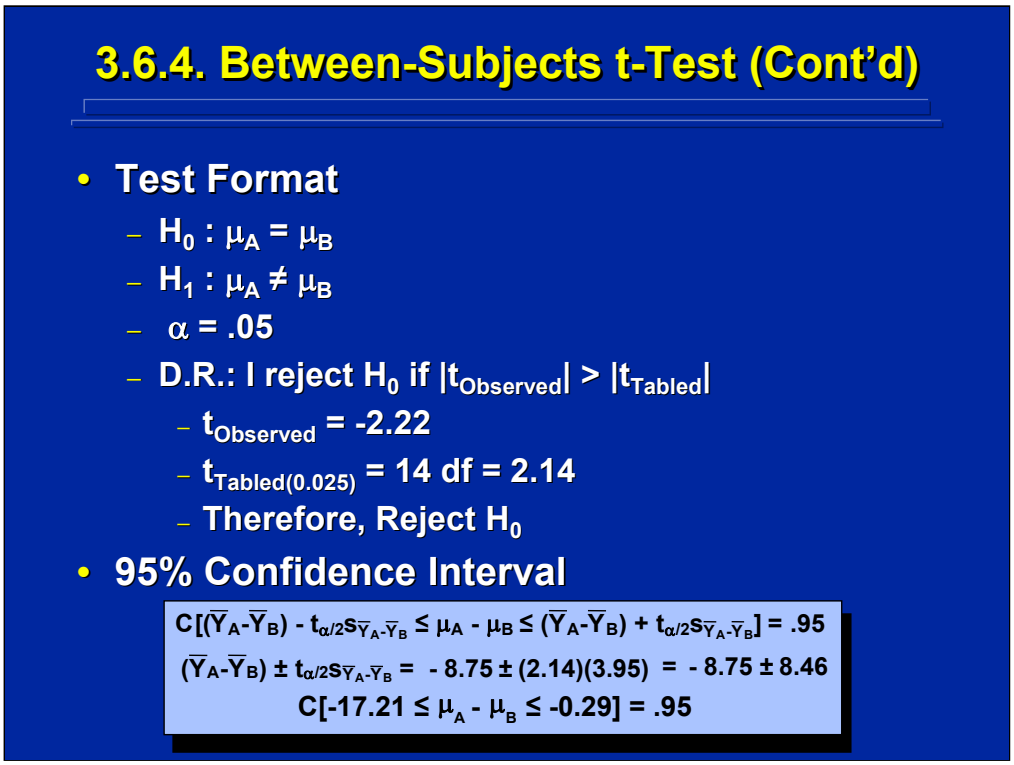
$$s^2_p = \frac{(7)(71.64) + (7)(52.86)}{8 + 8 - 2} = 62.25$$

$$t_{Observed} = \frac{\bar{Y}_A - \bar{Y}_B}{\sqrt{s^2_p(\frac{1}{n_A} + \frac{1}{n_B})}}$$

$$t_{Observed} = \frac{53.25 - 62.00}{\sqrt{62.25(\frac{1}{8} + \frac{1}{8})}} = \frac{-8.75}{3.94} = -2.22$$

(Click in this red rectangle to see SAS calculations for this example.)

The calculations of the t-observed using the pooled formula are shown on this slide. Alternatively, the definitional formula could have been used to calculate t-observed. The SAS program for conducting this between-subjects t-test is shown in Slater and Williges (2006) appendix.




(Click in this red rectangle to see SAS calculations for this example.)

The standard format for this example t-test is shown on the top portion of this slide. The null hypothesis states that the means are equal, and the two-tailed alternative states that they are not equal. Note that the tabled value is based on 14 degrees of freedom (i.e. 8 + 8 - 2), and $\alpha = 0.25$ in the t-table to make a two-tailed test of significance. The decision rule is to reject the null hypothesis if the t-observed value is greater than the t-tabled value. In this case the null hypothesis is rejected, and the experimenter concludes that using night vision display A resulted in significantly better squad maneuvering performance (i.e., lower mean maneuvering time) than using night vision display B.

The 95% confidence interval of the difference between the two means is shown on the bottom portion of this slide. Note that the range does not include zero, which would be the difference under the null hypothesis. Therefore, the null hypothesis is rejected just as in the statistical hypothesis test. All the hypothesis test tells the researcher is that the difference is statistically significant, not that it is practically significant. The experimenter must decide if an -8.75 minute statistically significant difference in mean squad maneuvering performance using the two displays is of any practical value in night vision display design.

## 3.6.5. Within-Subjects t-Test

- **Characteristics**
  - Same Subjects or Two Highly Correlated Samples
  - Homogeneity of Variance NOT an Issue
- $t_{Tabled\ (\alpha/2)}$ = (n - 1) df where, n = Number of Pairs
- $t_{Observed}$ Computational Formulae
  - Consider Relationship between Samples
    - Degree of Correlation, $r_{AB}$

$$r_{AB} = \frac{n\sum(Y_A)(Y_B) - (\sum Y_A)(\sum Y_B)}{\sqrt{\left[n\sum Y_A^2 - (\sum Y_A)^2\right]\left[n\sum Y_B^2 - (\sum Y_B)^2\right]}}$$

A within-subjects t-test is appropriate when the same subjects are observed either in both sample A and sample B or the subjects in each sample are closely matched on relevant characteristics. Consequently, the two samples being compared in the t-test are highly correlated. Within-subject t-tests also guarantee that the sample sizes are the same for the two samples. Consequently, there is no need to test the homogeneity of variance assumption. The t-tabled value has n-1 degrees of freedom, where n refers to the number of pairs of subjects or observations.

Since the two samples are related, it is necessary to consider the covariance between them when calculating the t-observed value from the data. The linear correlation coefficient, $r_{AB}$, is used to reflect the degree of relationship between the two samples. The raw score formula for calculating $r_{AB}$ is shown on this slide and this formula is described in detail in Topic 19. Note that it includes every observation from sample A (i.e., $Y_A$), and sample B (i.e., $Y_B$).

## 3.6.5. Within-Subjects t-Test (Cont'd)

- $t_{Observed}$ Computational Formulae (Cont'd)
  - Raw Score Formula

$$t_{Observed} = \frac{\bar{Y}_A - \bar{Y}_B}{s_{\bar{Y}_A - \bar{Y}_B}} = \frac{\bar{Y}_A - \bar{Y}_B}{\sqrt{\frac{s^2_{\bar{Y}_A}}{n_A} + \frac{s^2_{\bar{Y}_B}}{n_B} - 2r_{AB}\, s_{\bar{Y}_A} s_{\bar{Y}_B}}}$$

  - Difference Score Formula

$$t_{Observed} = \frac{\bar{D}}{s_{\bar{D}}} = \frac{\bar{D}}{\sqrt{\frac{s^2_D}{n}}} \qquad \text{where, } s^2_D = \frac{\sum_{i=1}^{n} (D - \bar{D})^2}{n - 1}$$

The t-observed value must be calculated either by using the raw score formula that includes the correlation between the two treatments as shown on this slide, or the difference score formula. The difference score formula directly incorporates the correlation between the two samples since only difference scores, D, and not the separate sample scores are used in the calculation as shown on the slide.

The two formulae for calculating t-observed are algebraically equivalent. However, the difference score formulae requires less calculation since $r_{AB}$ is not calculated, and the result is less prone to rounding errors resulting from the calculation of $r_{AB}$ in the raw score formula.

## 3.6.5. Within-Subjects t-Test (Cont'd)

- **Example Problem: An experimenter wishes to compare performance of two different night vision displays used in nighttime maneuvering. Eight squads used both Display A and Display B. Each squad completed the same nighttime maneuver twice. Half of the squads used Display A first and half used Display B first to counterbalance order of use. The experimenter wants to determine if there is a significant difference ($p < 0.05$) in mean time in minutes to complete the nighttime maneuver between using the two night vision displays.**

(Click in this red rectangle to see SAS calculations for this example.)

Assume the data presented in the between-subjects t-test example was collected from a within-subjects design. As described in this slide, the same 8 squads would use both night vision displays A and B. Consequently, a total of only 8 squads were needed to conduct this within-subjects experiment as compared to 16 different squads needed in the between-subjects design counterpart.

Since each squad completes the nighttime maneuver twice, the order of using the two night vision displays must be counterbalanced to avoid confounding display effects with practice on the nighttime maneuver. The easiest way to accomplish counterbalancing is to use an even number of squads such as 8 in this example. In this case, one can randomly select 4 of the squads to use night vision Display A first and the other 4 to use night vision Display B first.

## 3.6.5. Within-Subjects t-Test (Cont'd)

- **Hypothetical Within-Subjects Data Set**

| Night Vision Display A | Night Vision Display B | Difference Score (D) |
|---|---|---|
| $s_1 = 59$ | $s_1 = 54$ | $s_1 = 5$ |
| $s_2 = 65$ | $s_2 = 72$ | $s_2 = -7$ |
| $s_3 = 52$ | $s_3 = 69$ | $s_3 = -17$ |
| $s_4 = 45$ | $s_4 = 59$ | $s_4 = -14$ |
| $s_5 = 63$ | $s_5 = 67$ | $s_5 = -4$ |
| $s_6 = 42$ | $s_6 = 61$ | $s_6 = -19$ |
| $s_7 = 53$ | $s_7 = 51$ | $s_7 = 2$ |
| $s_8 = 47$ | $s_8 = 63$ | $s_8 = -16$ |
| $\bar{Y}_{PD} = 53.75$ | $\bar{Y}_{SD} = 62.00$ | $\bar{D} = -8.75$ |
| | | $s^2_D = 83.36$ |

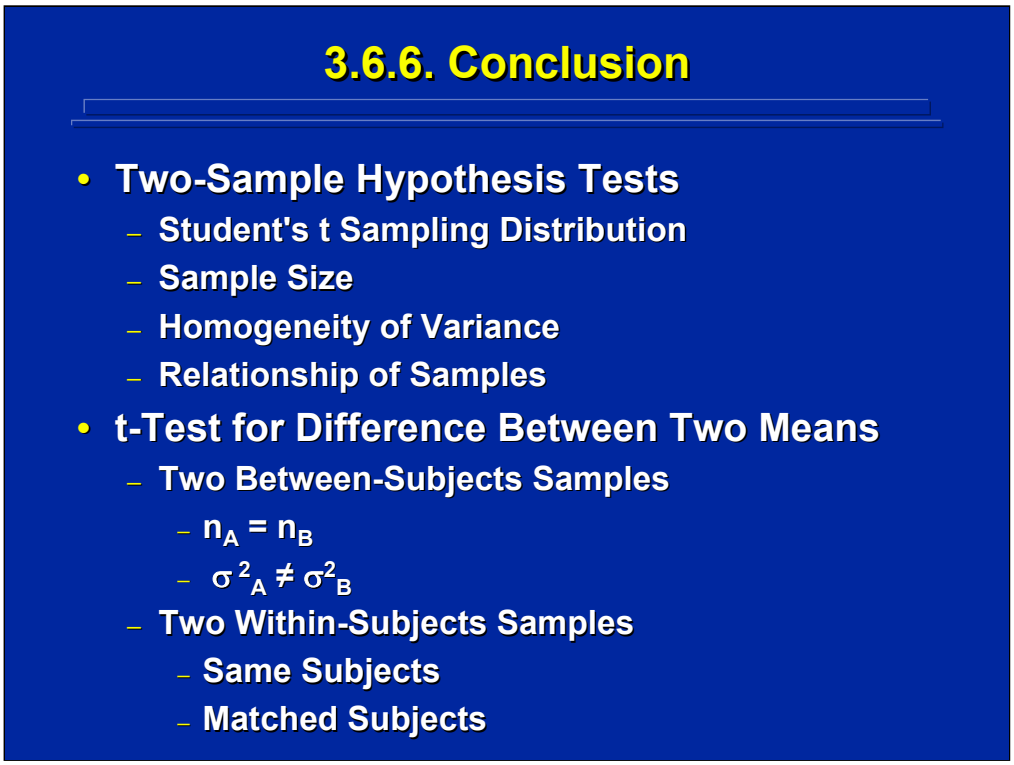
where, $t_{Observed} = \dfrac{\bar{D}}{\sqrt{\dfrac{s^2_D}{n}}} = \dfrac{-8.75}{\sqrt{\dfrac{83.36}{8}}}$

$t_{Observed} = -2.71$ $\quad$ $t_{Tabled(0.025)} = 7\ df = 2.36$

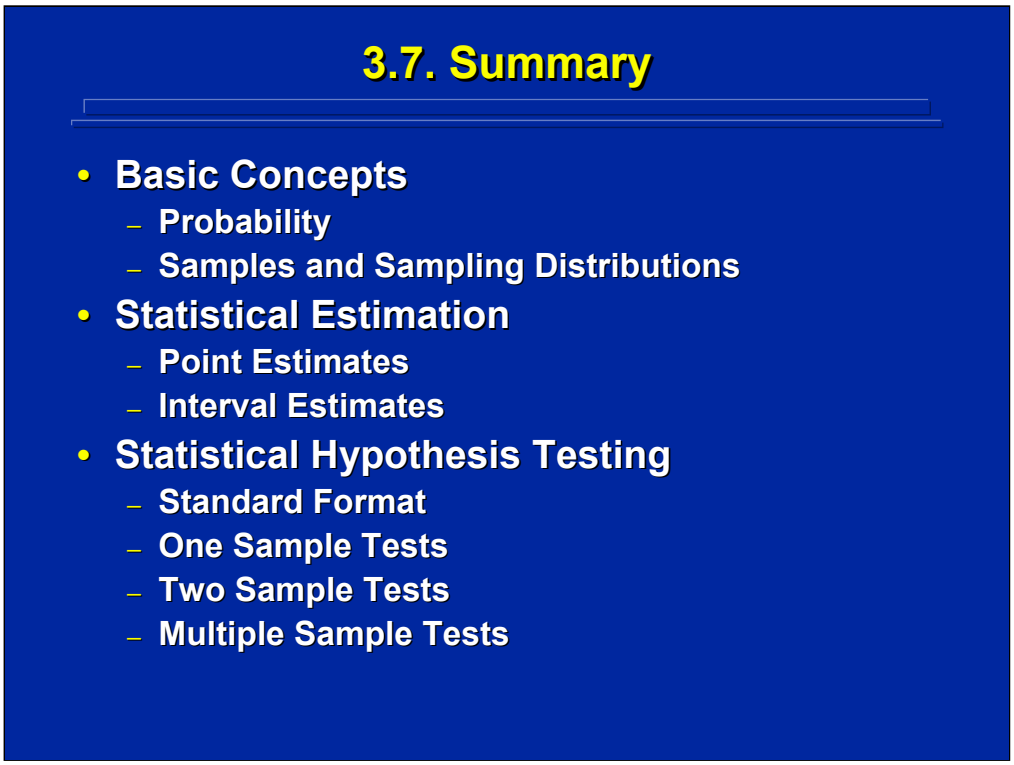
(Click in this red rectangle to see SAS calculations for this example.)

The difference score formula was used to calculate the observed value of the t statistic as shown on the bottom portion of this slide. The tabled value of t is based on 7 degrees of freedom which is equal to n-1, where n equals the number of different squads (i.e., 8). Since the absolute t-observed value is greater than the absolute t-tabled value, there is a significant difference between the squad mean maneuvering performance using the two displays. The SAS program for calculating this within-subjects t-test is shown in Slater and Williges (2006) appendix.

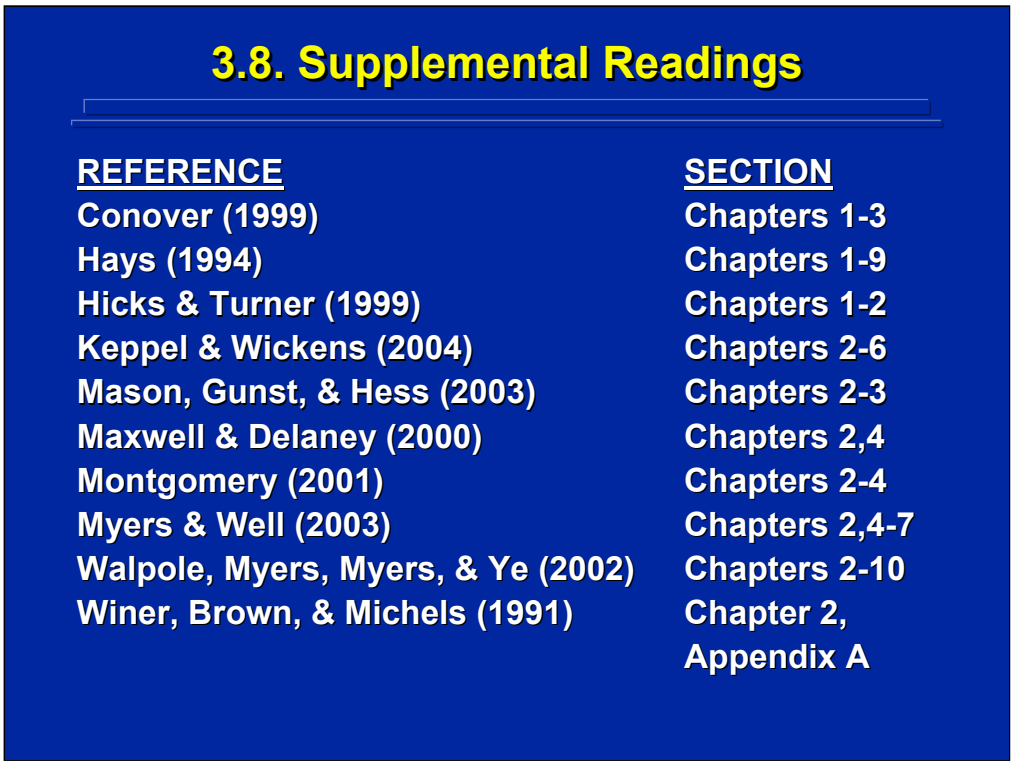
Even though the within-subjects t-test has a higher t-tabled value (i.e., lower degrees of freedom) than the previous between-subjects example (i.e., 2.36 versus 2.14), the results are still significantly different. The main effect of subject (i.e. squad) variability is removed from the within-subjects test. This reduction usually more than offsets the lower degrees of freedom, thereby making a within-subjects design alternative generally more sensitive then its between-subjects counterpart. Procedures for reducing subject variability are important in human factors research.




In conclusion, one normally uses the student's t test in human factors research due to small sample sizes that are less than thirty. If the samples are independent and sample size is equal, one does not need to consider homogeneity of variance unless the sample variances are markedly different. Calculation of the t-observed statistic depends on the relationship of the two samples. If the experimenter uses a between-subjects design, the samples are independent and homogeneity of variance should be tested if sample sizes are unequal. If the researcher uses a within-subjects design, the samples are related and the correlation between them must be considered. In both cases, the standard format for hypothesis testing is used.




Experimental design uses all the basic statistical concepts reviewed in this section. Probability, samples, and sampling distributions make up the fabric of experimental design analysis. Point estimates are the basic statistics used in experimental design, and interval estimation is directly related to hypothesis testing. Statistical hypothesis testing is the primary inferential process used in experimental design. Every hypothesis test can be stated in a standard format that includes the null hypothesis, alternative hypothesis, alpha level, and decision rule. This section reviewed one-sample and two-sample hypothesis testing. Two sample tests are the simplest designs used in one-way, between-subjects and within-subjects ANOVA. Most human factors research problems are more complicated and require multiple sample tests that are considered in basic ANOVA.




Several texts dealing with basic statistical concepts are listed on this slide for supplemental reading and a more detailed discussion of the basic statistical concepts reviewed in this topic. The Walpole, Myers, Myers, and Ye (2002) is a comprehensive introductory statistics text written for scientists and engineers and provides a good mathematical treatment of the topic. Hays (1994) is a classic behavioral science statistics text that relates the basic statistical concepts to research using human subjects. Finally, Conover (1999) provides a discussion of probability theory, the discrete binomial sampling distribution, and the use of the binomial distribution in hypothesis testing.

The rest of the references shown on this slide are experimental design texts that will be referenced throughout this reference material. The appropriate chapters in these texts that review basic statistical concepts are listed for each text. Of these, the Keppel and Wickens (2004), Myers and Well (2003) and the Winer, Brown, and Michels (1991) books are classic experimental design texts in the behavioral sciences and provide both a conceptual and mathematical treatment of many of the basic statistical concepts reviewed in this topic.

# Section 2.
# Supplemental Data Collection and Analysis

**Topic 4. Supplemental Data Collection Methods**
**Topic 5. Analysis of Nominal Scale Data**
**Topic 6. Analysis of Ordinal Scale Data**
**Topic 7. Summary of Supplemental Data**

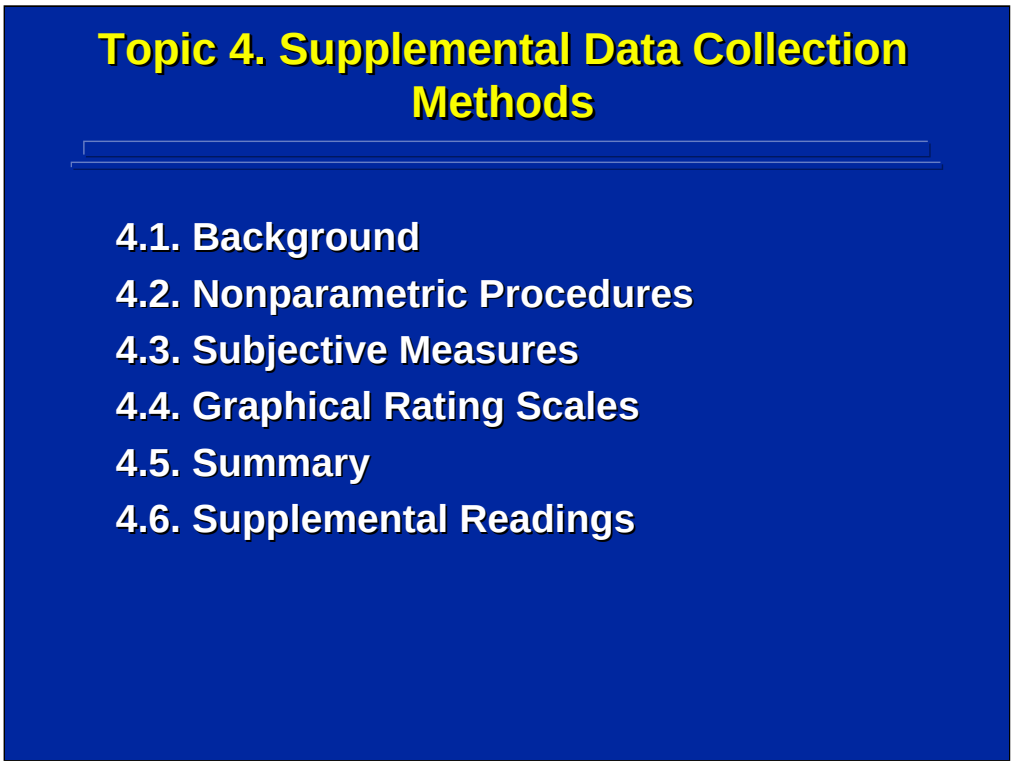
All supplemental data collection and analyses are designed to provide a richer interpretation of the results of primary performance data analyses from an experimental design. Before embarking on a discussion of basic and complex ANOVA designs, this section of the reference material provides:

Topic 4 - an overview of methods used to collect supplemental data;

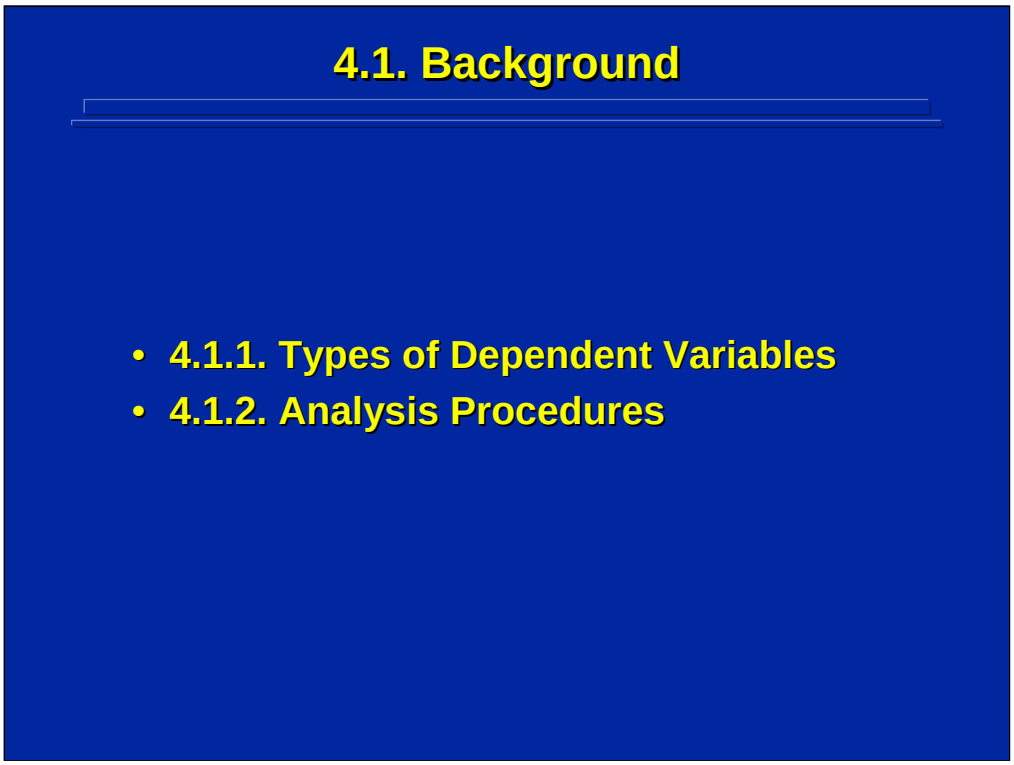
Topic 5 - basic nonparametric methods for analyzing supplemental nominal data that is in the form of frequency counts; and

Topic 6 - basic nonparametric methods for analyzing supplemental ordinal data that is in the form of rank orders.

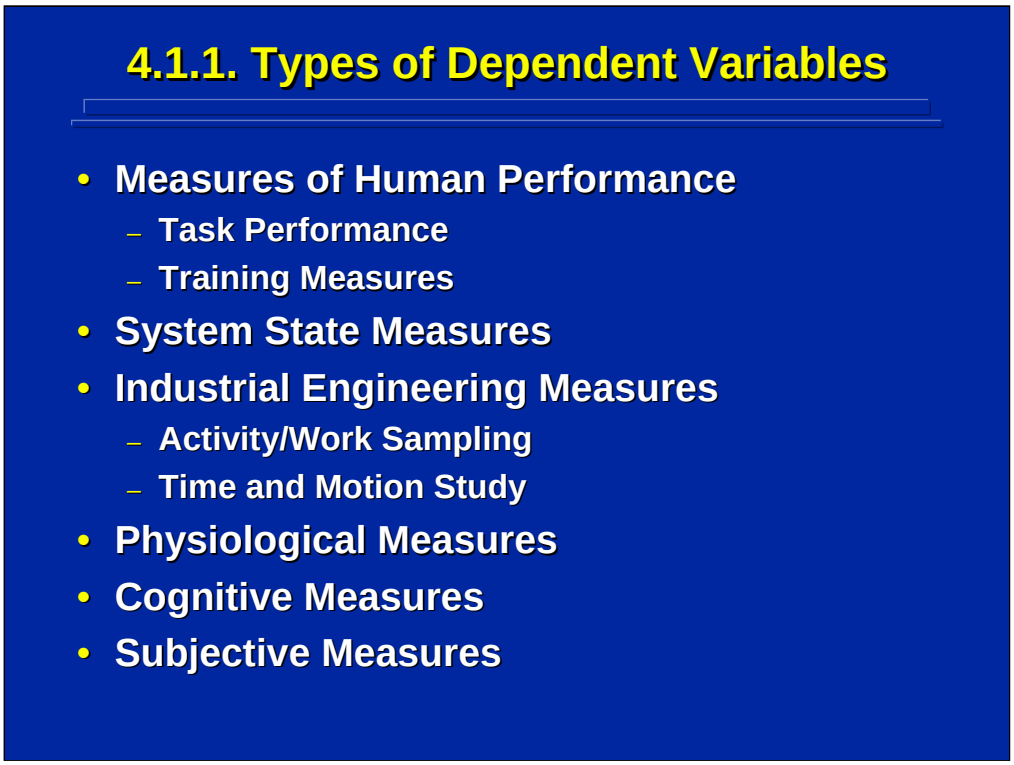
Topic 7 – summary and process for dealing with supplemental data



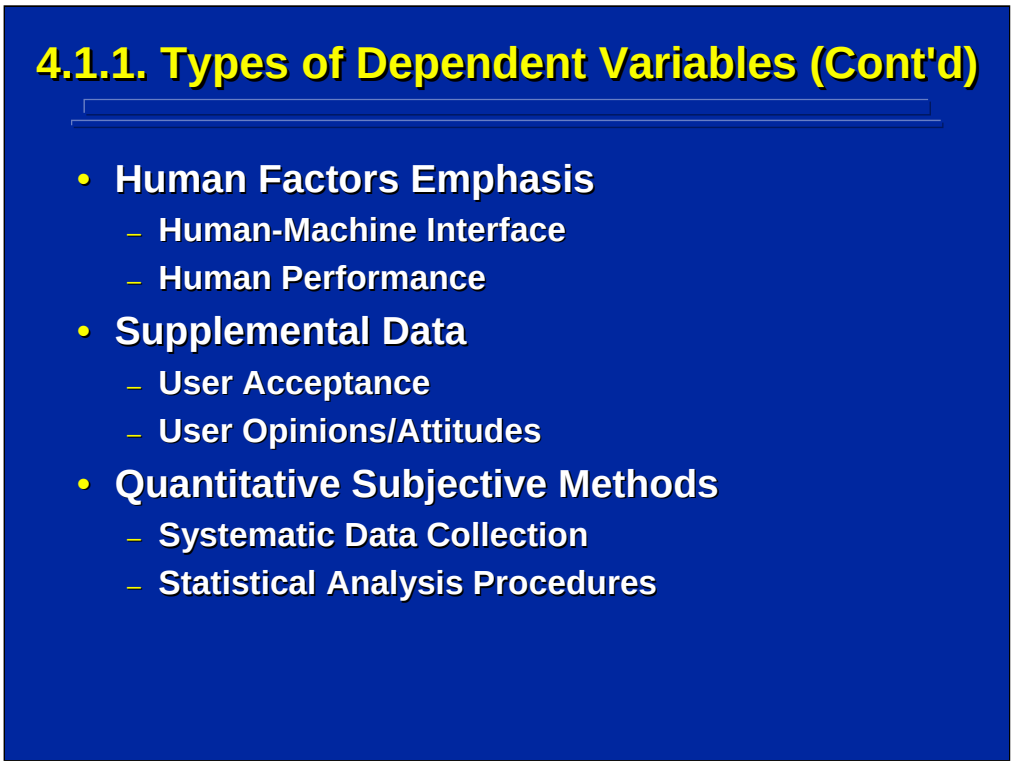

This topic deals with supplemental data collection that augments data from experimental designs. Supplemental data are very important for understanding and interpreting research results. Try to make the supplemental data as quantitative as possible, because it will be easier to analyze and incorporate into the results. Most supplemental data are analyzed by nonparametric analyses as opposed to parametric analyses used on the major dependent variables manipulated in the experimental design. Several types of supplemental data collection techniques can be used. This reference will concentrate on graphical rating scales that are most often used to collect quantitative supplemental data in human factors research.



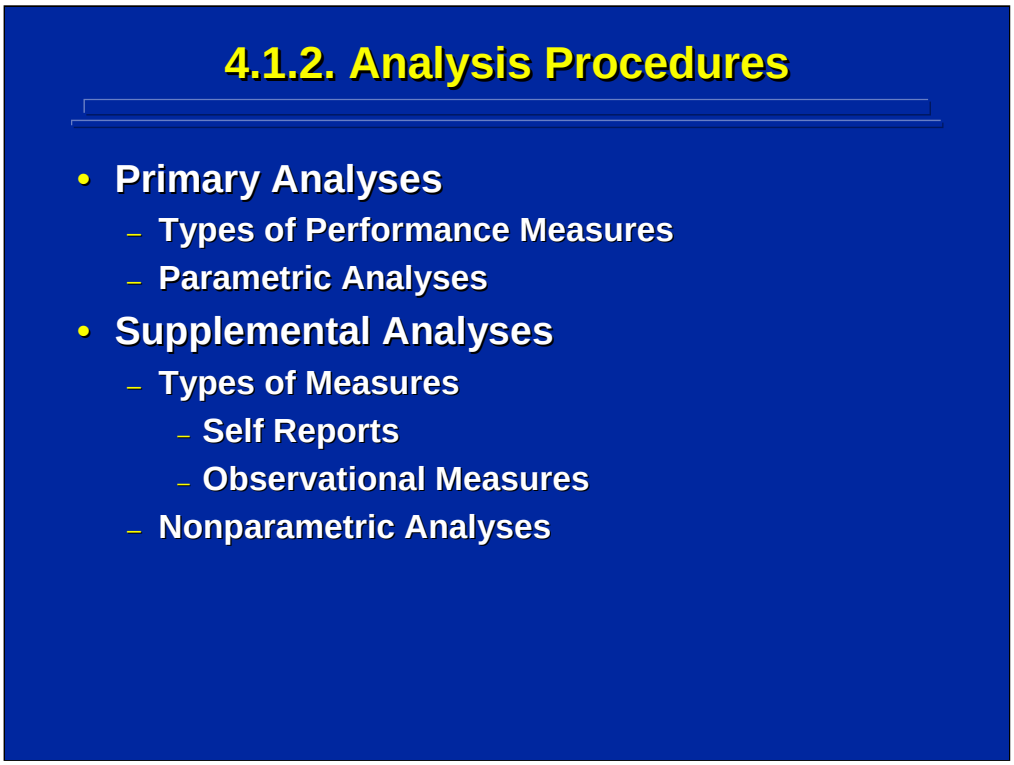

Before discussing rating scales in detail, one needs to consider the types of dependent variables and the analysis procedures used with supplemental data.



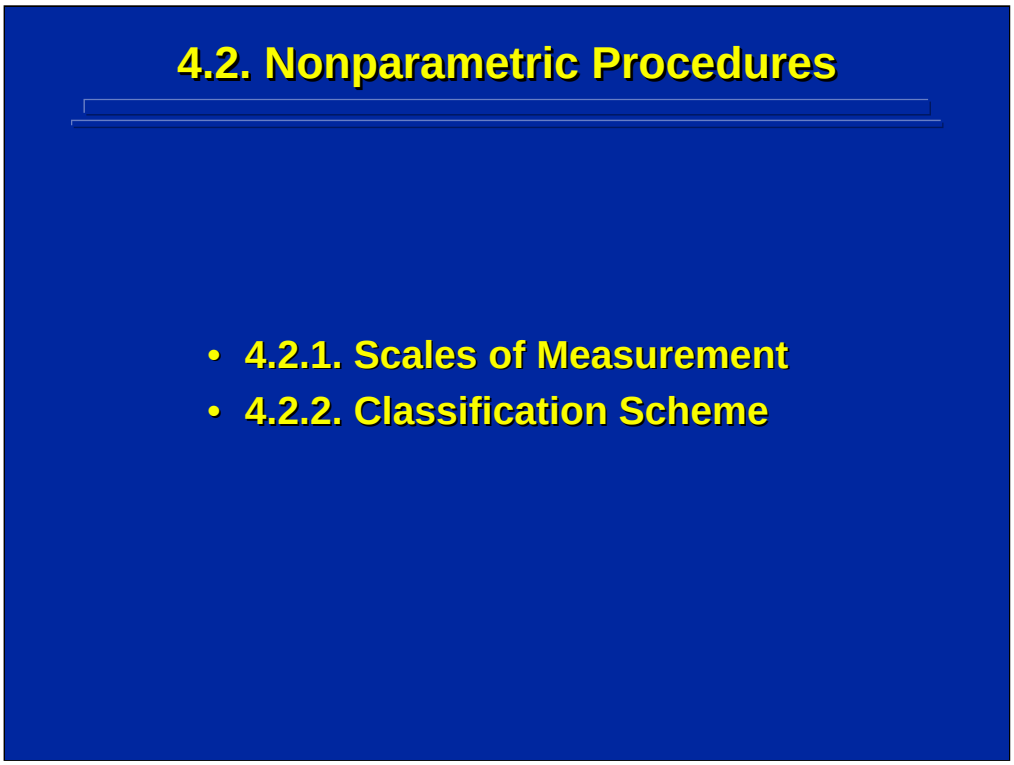

This slide shows the variety of dependent measure of interest to human factors researchers. Experimental designs usually investigate human performance metrics, but they can also evaluate measures of system states, industrial engineering measures of work activity, physiological, and cognitive metrics. Most supplemental data collection and analysis deal with subjective measures that are collected as objectively as possible.




Human factors research is focused primarily on human-machine interface design and human performance evaluation. Both user acceptance and user opinions are key aspects of supplemental data to augment performance evaluation. These data are subjective in the sense that the human subjects are requested to provide data in the form of opinions, satisfaction, suggestions, etc. in addition to their primary task performance in the experiment. However, researchers want these subjective measures to be collected systematically and to be as quantitative as possible so that statistical analysis procedures can be used to analyze the results.




This slide characterizes the difference between primary and supplemental data analyses. Experimental designs provide the data for the primary analyses. These analyses use parametric techniques such as ANOVA to test statistical hypotheses based on interval data. On the other hand, supplemental analyses use primarily nonparametric techniques based on frequency counts and rank orderings based on observations and self-reports.



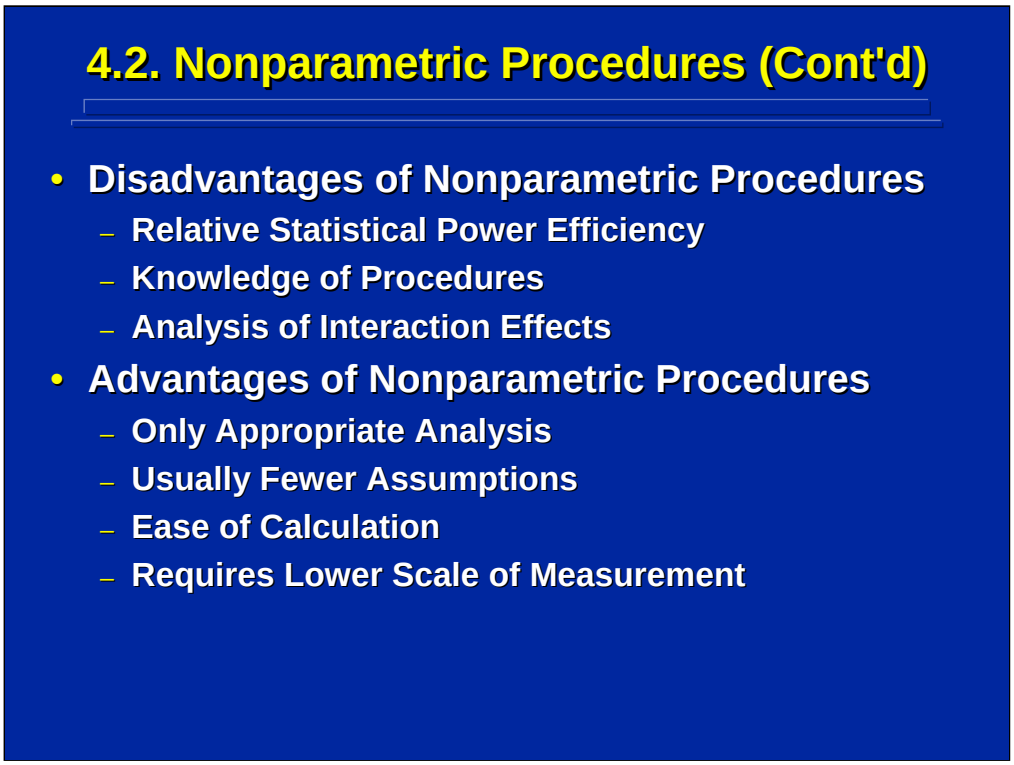

Nonparametric analyses depend upon the scale of measurement that is present in the supplemental data. Characteristics of various scales of measurement are reviewed, and a classification scheme for alternative nonparametric analyses based on this measurement scale is presented.

## 4.2. Nonparametric Procedures (Cont'd)

- **Definition: Nonparametric statistics do not have underlying assumptions (e.g., normal distribution, equal variance) and use mathematical procedures appropriate for nominal and ordinal data.**
- **Parametric vs. Nonparametric Procedures**
  - **Underlying Assumptions**
  - **Scale of Measurement**
- **Power Efficiency of Statistical Tests**
  - **Definition: Power efficiency is the required increase in sample size of Test B, $N_B$, to make it as powerful as Test A when the sample size of Test A, $N_A$, and the level of significance is held constant.**

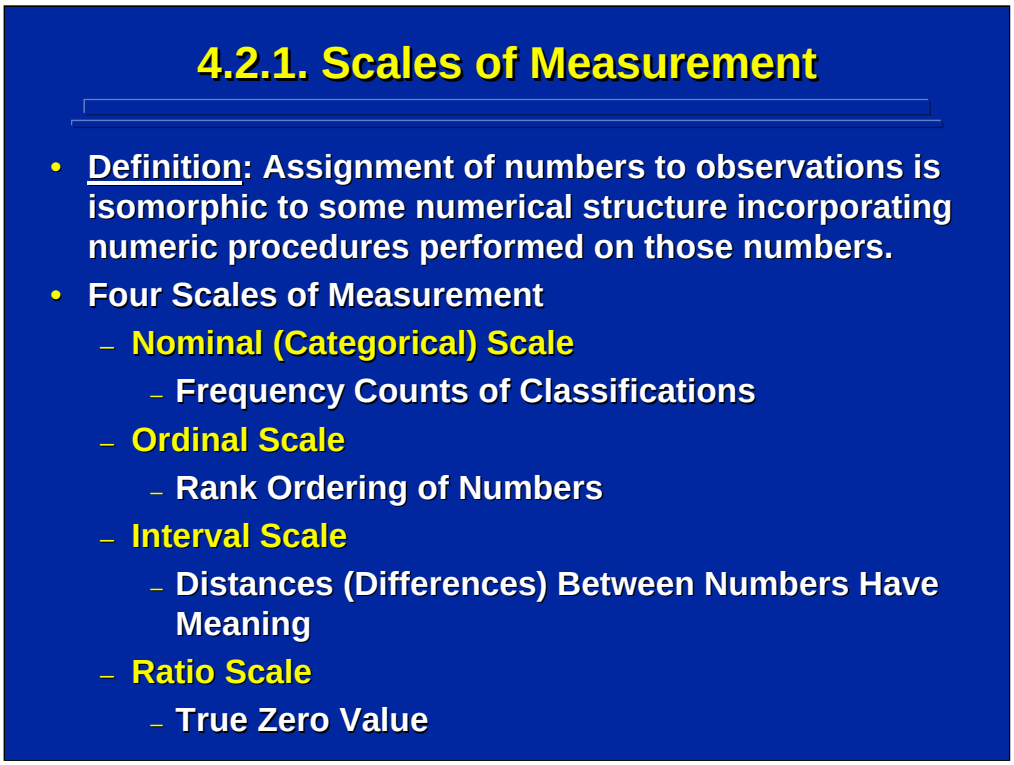
Parametric statistics have basic assumptions and perform statistical tests on parameters using numeric procedures appropriate for interval or ratio scale qualities. Many of the supplemental data do not have these qualities. So, the experimenter must use nonparametric analysis. The two basic differences between parametric and nonparametric procedures are the underlying assumptions and the scale of measurement underlying the data.

Since nonparametric analyses do not have the assumptions and do not have the numeric characteristics of data used in parametric analyses, they are not as powerful as their parametric analysis counterpart. Power efficiency shows this effect by stating the percent increase in sample size needed to make a nonparametric test as powerful as a parametric test for the same level of significance.

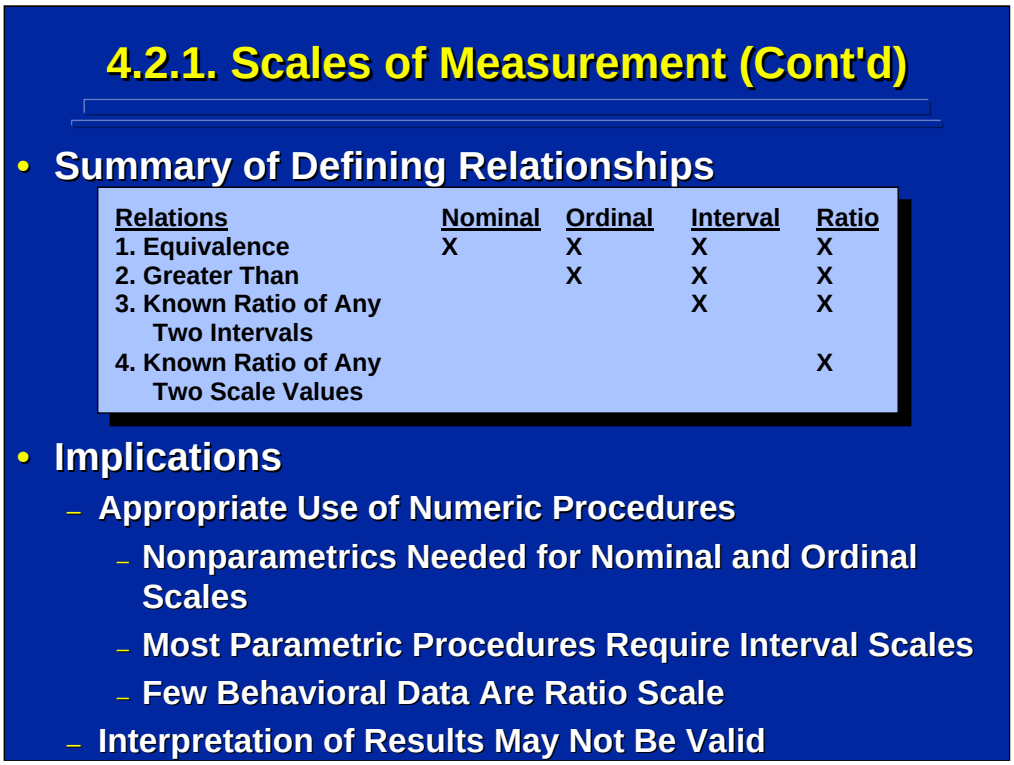



The major advantages and disadvantages of nonparametric analysis are listed on this slide. The primary disadvantage of a nonparametric test is that it has lower power efficiency. In addition, many researchers do not know the various nonparametric techniques and when to use them. Finally, it is often difficult to analyze interactions directly in a nonparametric analysis requiring additional subsequent analysis to isolate significant interaction effects.

On the other hand, nonparametric procedures offer several advantages. Sometimes a nonparametric test is the only appropriate test for supplemental data. Usually fewer assumptions are required to conduct a valid nonparametric analysis. Most nonparametric procedures are easy to calculate manually in lieu of using computerized statistical packages. Finally, nonparametric analyses are designed to analyze frequency counts or rank order data that often compromise supplemental data instead of the interval data needed for parametric analysis.

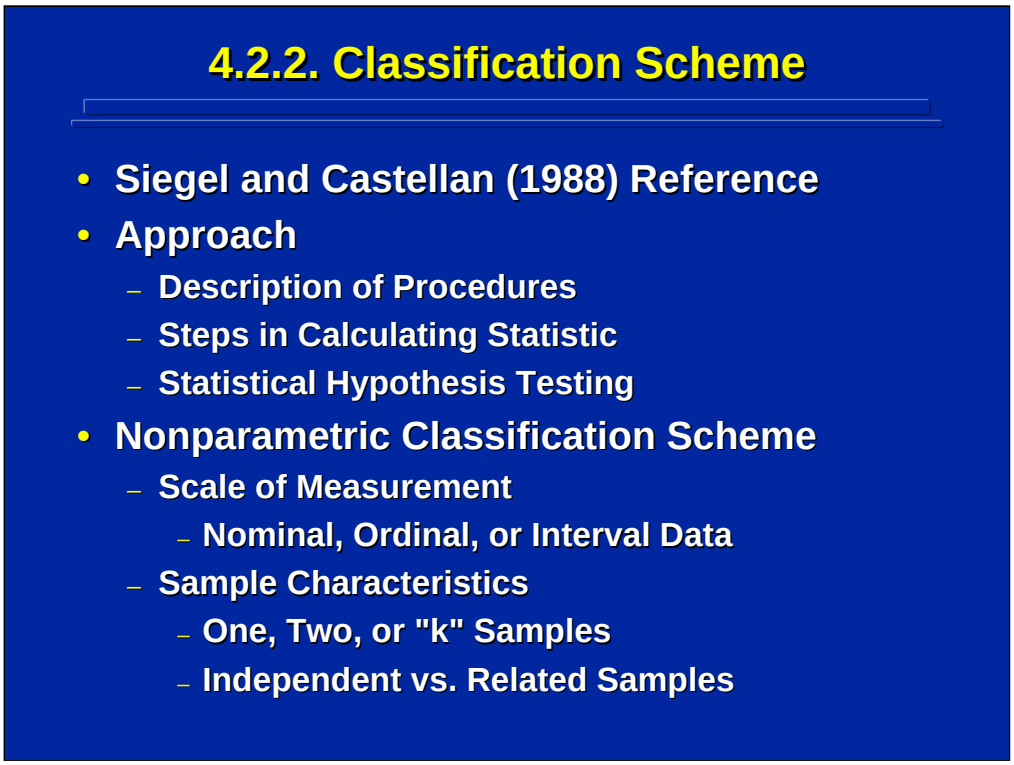



This slide provides a definition of a measurement scale. There are four scales of measurements according to Stevens (1951). A nominal scale or categorical scale involves just frequency counts of classifications. An ordinal scale is the rank ordering of numbers across intervals, but the intervals are not necessarily equal. Interval scale exists when the distances or differences between intervals have meaning. Finally, a ratio scale has the characteristics of all the other scales in addition to a true zero value.

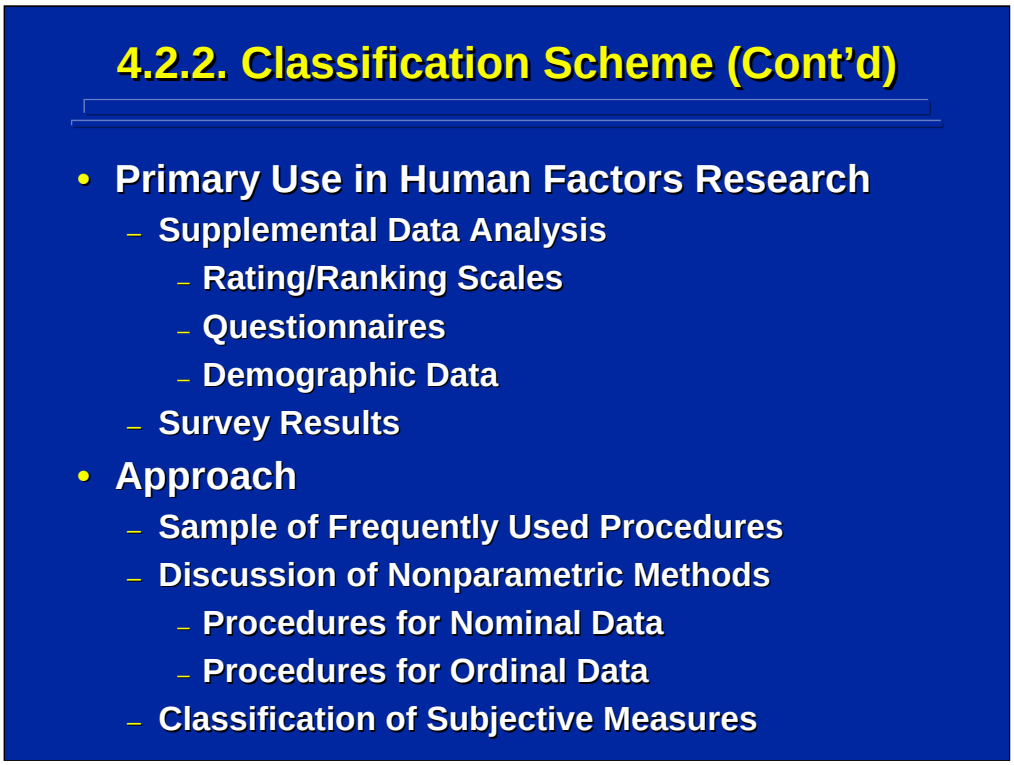



The top portion of this slide summarizes the characteristics of the various scales of measurement in ascending order. Parametric analysis requires at least an interval scale. So, nonparametric analysis is needed for nominal and ordinal scale data.

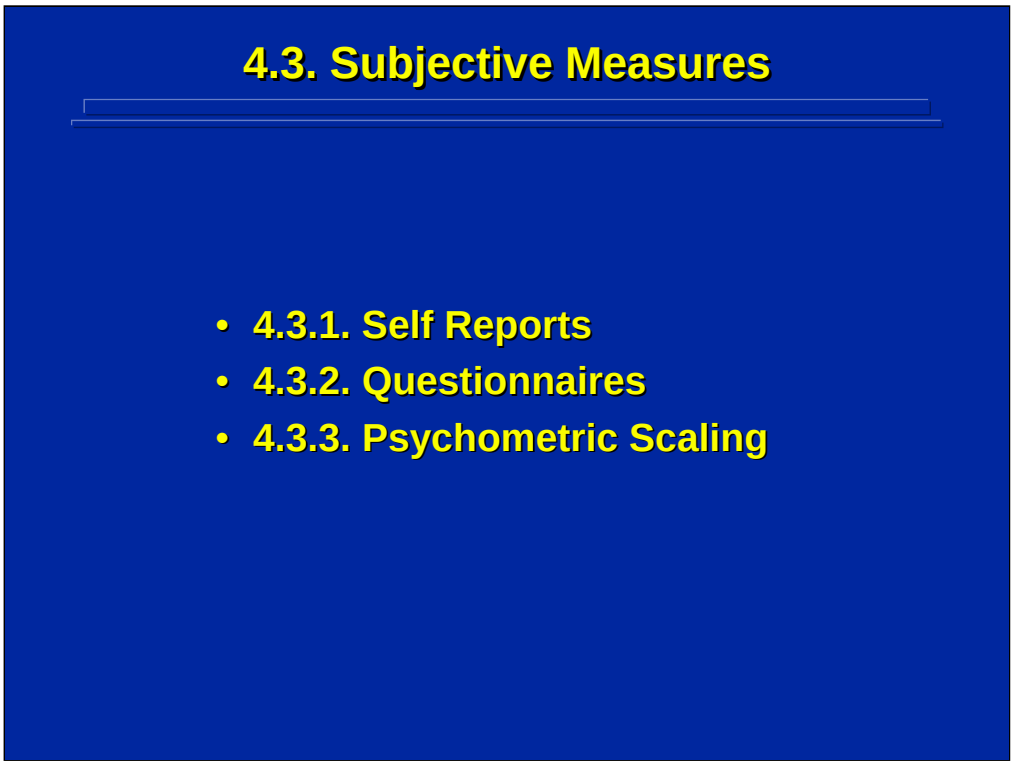
The major implication of different measurement scales is interpretation, not analysis. If the data do not exhibit the characteristics of the measurement scale used in the analysis, then the interpretation may not be valid. Often the choice of analysis is straightforward in human factors research. For example, measures such as accuracy, speed and time are evaluated by a parametric analysis because interval scale interpretations are made. If, on the other hand, the data just exist as frequency counts or rank orders, then the experimenter should consider using a nonparametric analysis for valid interpretation. Sometimes the choice is not straightforward. For example, rating scale evaluations use either parametric or nonparametric analyses depending on the assumed underlying qualities of ordinal and interval characteristics that are built into the scale.




The Siegel and Castellan (1988) approach and classification scheme for nonparametric analysis is often used by human factors researchers. They describe the nonparametric procedure, then the steps in calculating the statistic, and finally the procedure for hypothesis testing. Their discussion is classified by the scale of measurement, the number of samples in the data set, and independent or related sample relationships (i.e., between-subjects or within-subjects experiments).




In human factors research, nonparametric analyses are primarily used for analyzing supplemental data that are in the form of rating/ranking scales, questionnaires, demographic data, or survey results. This reference material reviews a sample of some of the frequently used procedures for nominal and ordinal subjective measures.




There are several ways to generate subjective measures that are quantitative. The three alternatives shown on this slide (i.e., self reports, questionnaires, and psychometric scaling) are used quite often in human performance and cognitive research. Important considerations for each alternative are described separately.

## 4.3. Subjective Measures (Cont'd)

- Subjective Data: Observations and Opinions
- Classification of Observations
  - Self Observations
  - Observation of Others
  - Observation of Events
- Qualitative vs. Quantitative Methods
- Goal for Analysis: Quantify
  - Frequency Counts
  - Psychometric Procedures
- Objective Analysis of Subjective Data
  - Avoid Subjective Analysis
  - Avoid Bias by Careful Design

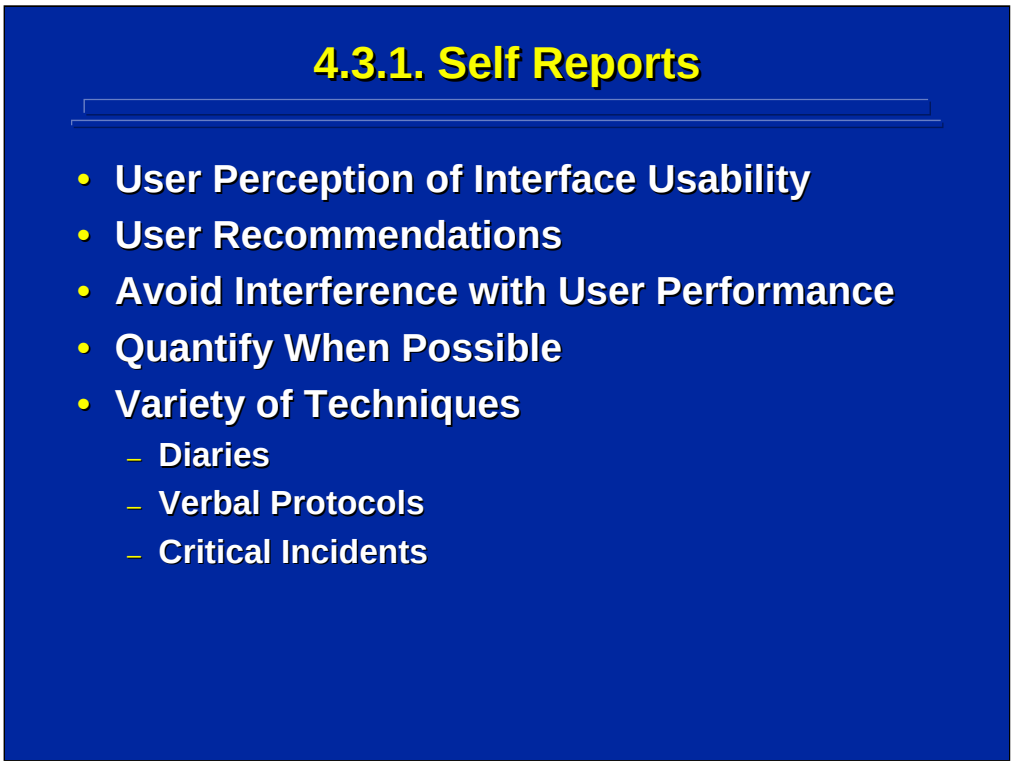
Subjective data are observations and opinions used in supplemental analyses related to an experiment. Observational data can be classified as self-observations made by the individual participating in the experiment, observations made by the experimenter, or observations of events occurring during the experiment.

The key consideration is to make the subjective data as objective as possible by using standard data collection procedures that result in quantifiable results. Just because the subject in the experiment or the experimenter generates the data subjectively does not mean that the subsequent analysis cannot be objective. Be careful not to make a subjective analysis of subjective data.

## 4.3. Subjective Measures (Cont'd)

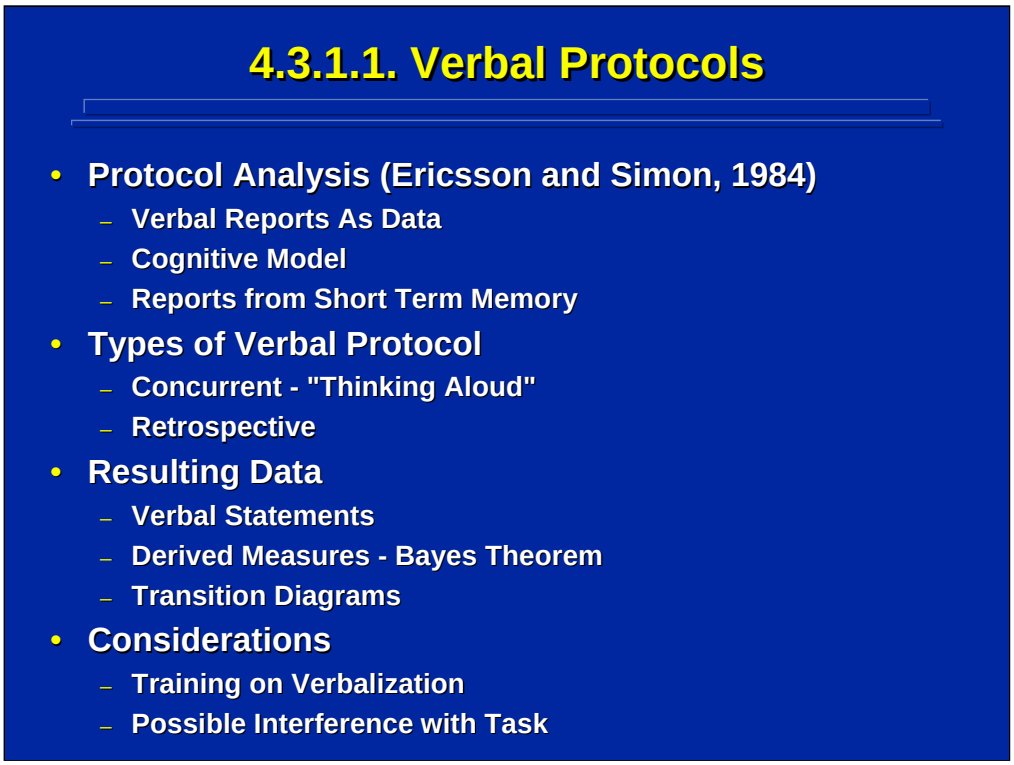
- **Examples of Inappropriate Subjective Analysis (Pew, 1993)**
  - **1. Display designers represent users during evaluation.**
  - **2. Final test and evaluation of a system is based solely on verbal protocol data.**
  - **3. Opinions are solicited without the opportunity to experience evaluation conditions.**
  - **4. After test pilots fly a new display, they exchange views before making a single recommendation.**

This slide lists four common examples of subjective analysis described by Pew (1993). These examples underscore that subjective data must come from the actual user rather than the designer; the final test and evaluation of a system should not be based solely on subjective data; the user must have an opportunity to experience the conditions to be evaluated before providing opinions; and care must be taken to collect the subjective data independently for each subject.



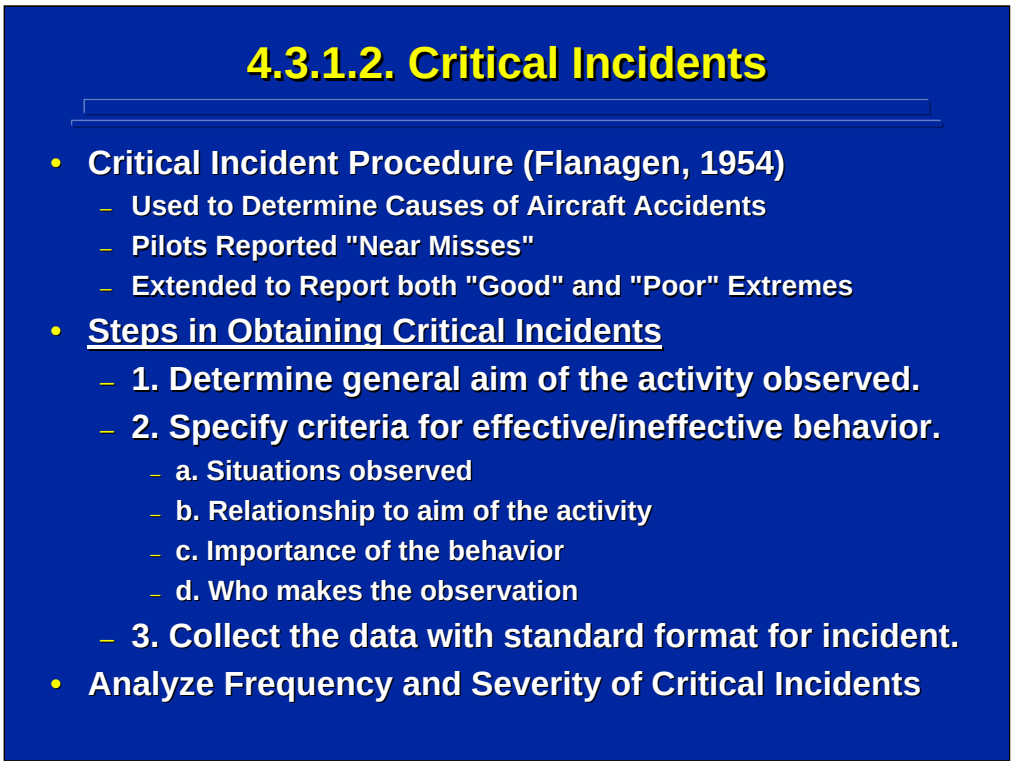

Self reports are user perceptions of the interface and their recommendations for improvements. Care must be taken to insure that user self-reports are gathered in a way to avoid interfering with actual task performance, and self-reports should be quantified whenever possible.

The bottom of this slide lists three common approaches to collecting self-reports in human factors research. A straightforward and informal way of collecting self-report data is to require participants to keep a diary of their perceptions throughout the experiment. Verbal protocols and critical incidents, however, are formal procedures for collecting self-report data and are discussed separately in this reference.



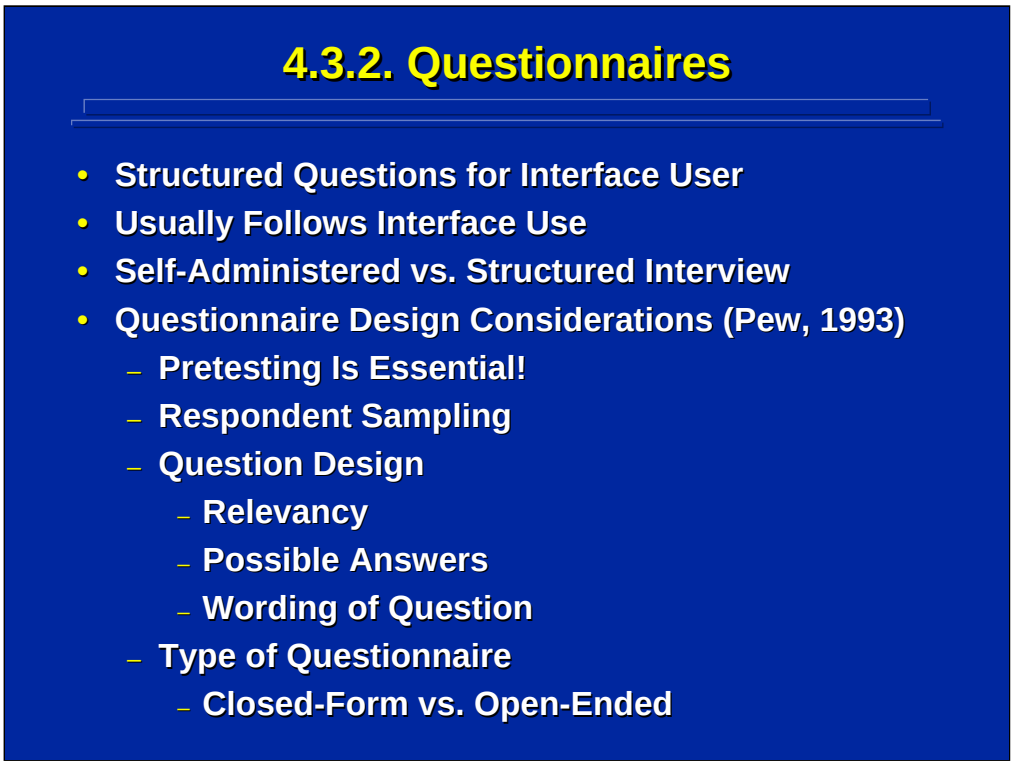

Ericsson and Simon (1984) described a technique called verbal protocol analysis in which verbal reports can be viewed as data that are recalled from the user's short-term memory. Their technique is often used in human factors research especially in human-computer interface design. Verbal protocol data can be collected concurrently while subjects perform a task or retrospectively after the task is completed possibly by viewing a videotape of their performance and describing what they were doing during each step of the task. The major problem with concurrent verbal protocols is interference with primary task performance, and the major problem with retrospective verbal protocols is forgetting. The resulting data are in the form of a verbal statement from the person. Many times the subsequent analysis is no more than sorting the information into meaningful categories and analyzing the frequency of response across categories. Ericsson and Simon (1984) also described derived measures based on Bayes Theorem and the use of transition diagrams as a more formal way of analyzing verbal protocols.

There are two primary considerations that an experimenter must address in using verbal protocols for self-reports. First, the experimenter must provide some training for the subject on verbalization. Second, the researcher must guard against possible interference with the task being performed.




Flanagen (1954) described a critical incident method of self-report developed by human factors specialists as a way to investigate causes of aircraft accidents. This method was used to obtain self reports from pilots after experiencing critical incidents during flying that could have resulted in catastrophic accidents. This procedure can be extended to look at the extremes of both the best and worst performance in order to find the good design aspects to keep and the bad design aspects that should be eliminated.

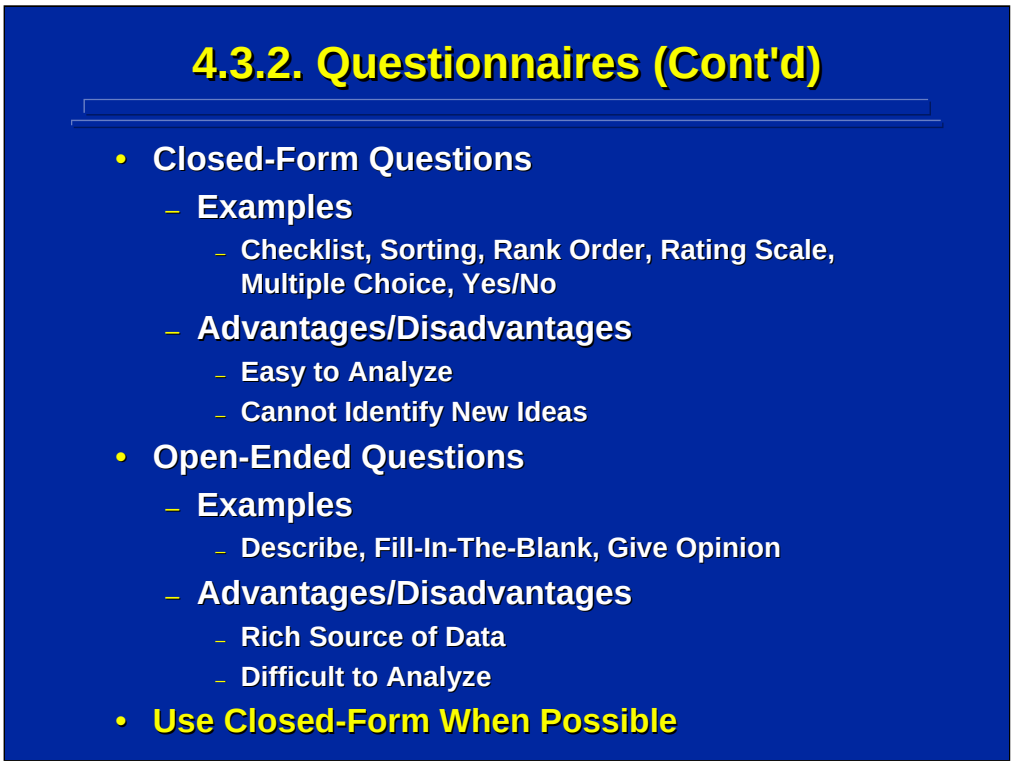
The middle portion of this slide summarizes the steps taken to obtain critical incidents. Collecting data in a standard format facilitates subsequent analysis of the self report. The resulting critical incidents are grouped into homogeneous categories and analyzed in terms of both frequency and perceived severity of the incident.




The most common self-report procedure used in collecting supplemental data is a questionnaire. The questionnaire is usually presented after the task is completed, but sometimes it is presented before an experiment to collect demographic data during subject selection. Questionnaires can be self-administered with a completion form or the data can be collected through a structured interview.

Pew (1993) discussed three major design aspects of questionnaires that need careful consideration when using them for supplemental data in experiments. Pretesting the questions is essential! Questionnaires usually need to be revised in order to provide the proper coverage. Poorly designed questions in terms of relevance, inappropriate answers, and ambiguous wording can yield ambiguous and unreliable results.

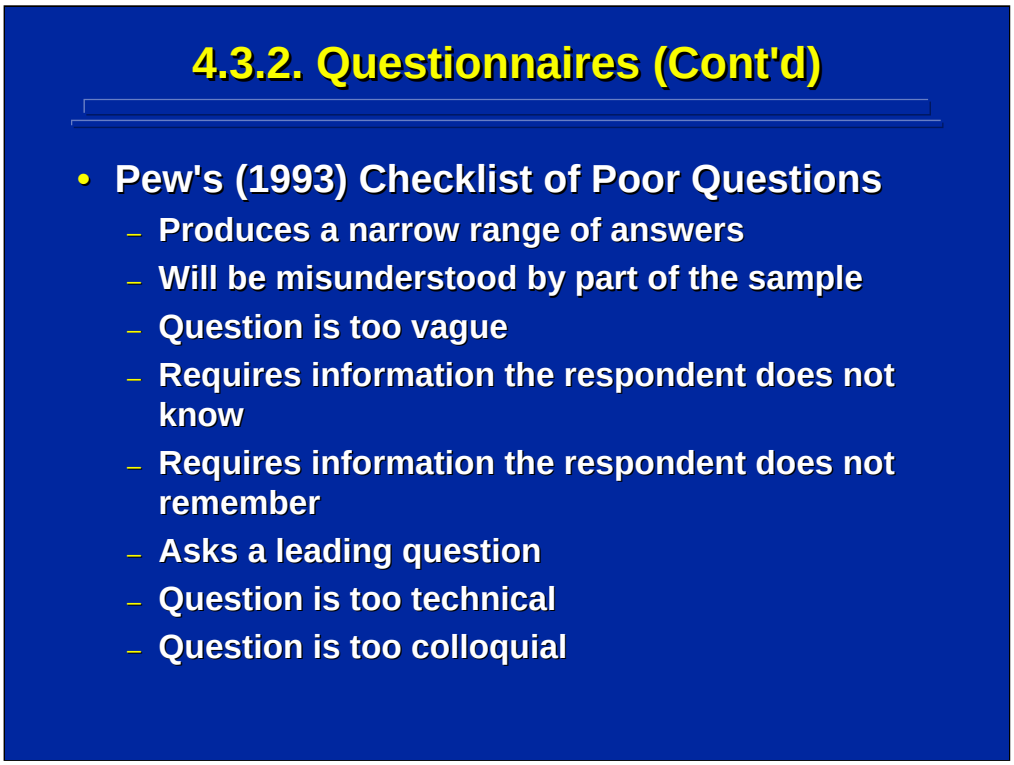
There are two types of questions, closed-form or open-ended. Closed-form questions provide a structured choice of possible answers; whereas, open-ended questions do not restrict the possible answer alternatives. Both types of questions can be used in collecting supplemental data.



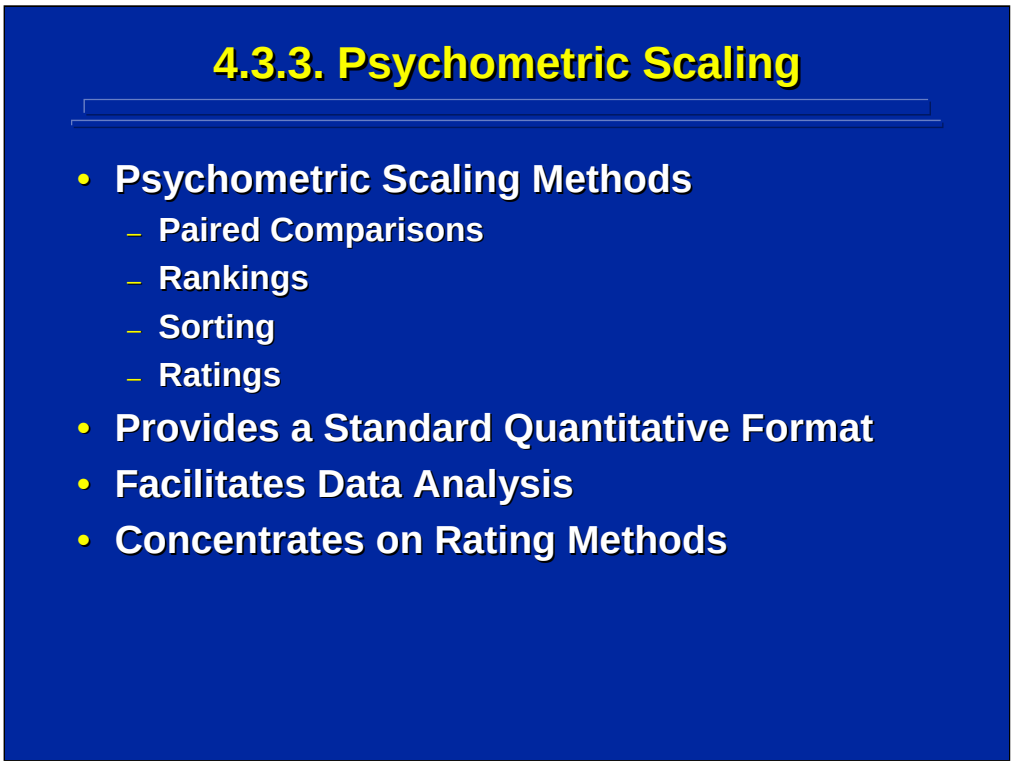

Closed-form questions are exemplified by checklists, rank ordering, sorting, yes/no answers, multiple choice answers, and rating scales. They provide a pre-specified choice of specific answers that facilitate analysis, but they do not allow for the expression of new ideas.

Open-ended questions are characterized by making general descriptions, fill in the blank, or stating opinions. Open-ended questions are unstructured and request general views that may be more difficult to summarize but yield a broader range of responses to gain new insights and ideas. Open-ended questions can be conditional when combined with rating scales by asking the user for clarification of only extreme rating responses.

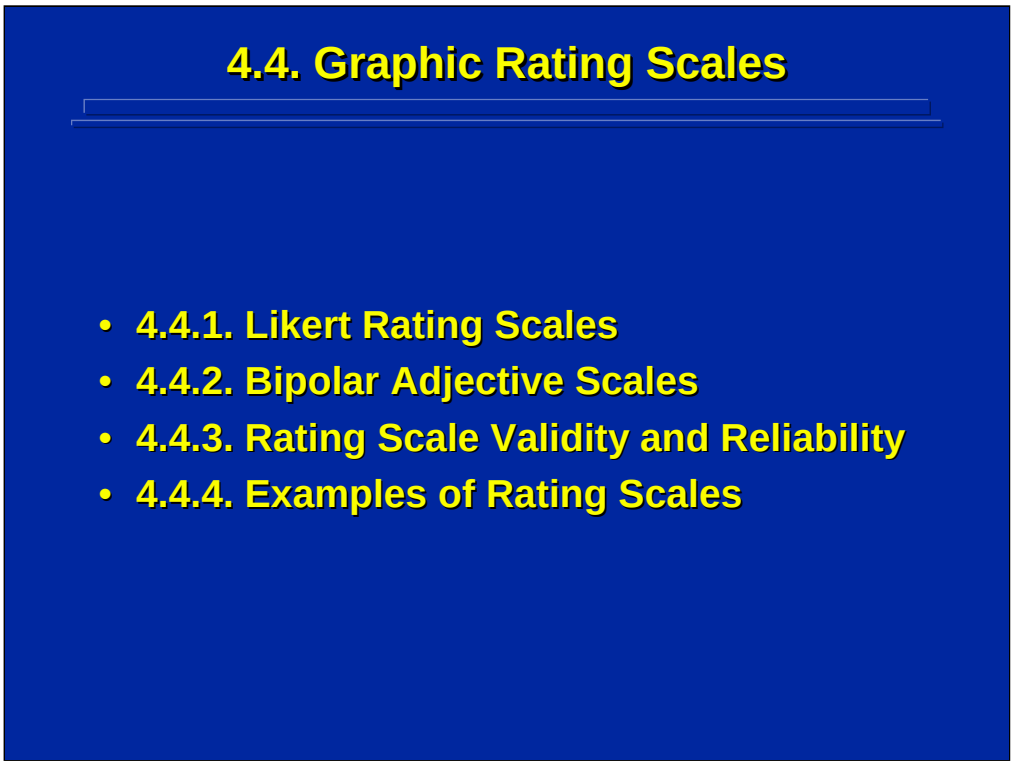
Usually a questionnaire designed for supplemental data collection is closed-form and explores specific issues related to the experiment. Often one final open-ended question is provided to obtain the participants' overall impressions and suggestions. To facilitate subsequent analysis, closed-form questions that have been carefully designed and pretested are preferred.




This slide summarizes Pew's (1993) checklist of poor questions that can be used in questionnaire design and pretesting.




Psychometric scaling methods include paired comparisons, rankings, sorting, and ratings. They use closed-form questions, provide a quantitative format, and require measurement properties that often go beyond nominal scale metrics. These techniques facilitate subsequent analysis of self-reports. Rating methods are the most common psychometric scales used in human factors research. This reference concentrates on some specific rating methods that are often used in human factors research.



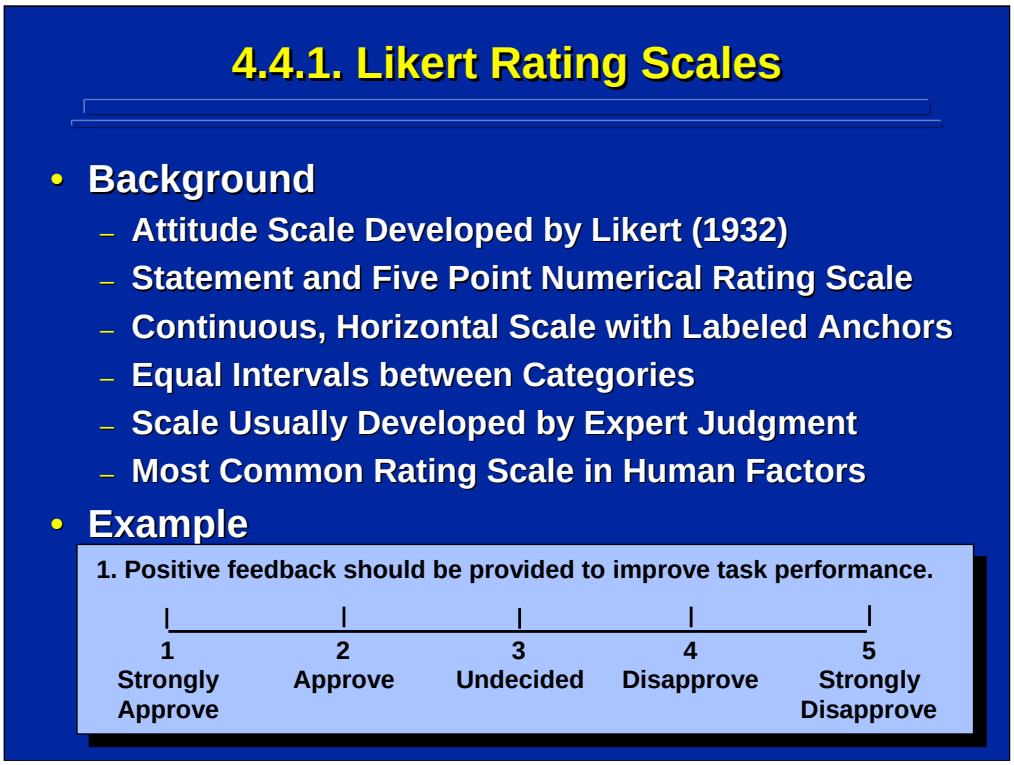

This subsection provides an overview of two of the most common types of rating scales used in human factors and ergonomics research, Likert scales and bipolar adjective scales. Validity and reliability are important issues in rating scale construction. This subsection ends with two examples of rating scales used for supplemental data collection in support of experimental design.

## 4.4. Graphic Rating Scales (Cont'd)

- User Sentiments/Attitudes/Opinions
- Definition: Unbroken line or boxes with labeled divisions representing a characteristic, behavior, or dimension to be rated.
- Variations in Graphic Rating Scales
  - Scale Orientation - Vertical or Horizontal
  - Number of Categories - Usually 5 to 9
  - Order of Scale - Positive or Negative
  - Center Point - Present or Absent
  - Labels (Anchors) - Words and/or Numbers
- Various Scale Development Procedures

Rating scales are the most popular way to collect supplemental data for human factors experiments because the results yield numbers that are amenable to quantitative analysis. Ratings can be used to measure a subject's sentiments, attitudes, or opinions. Most scales used in human factors and ergonomics research are some form of a graphical rating scale.

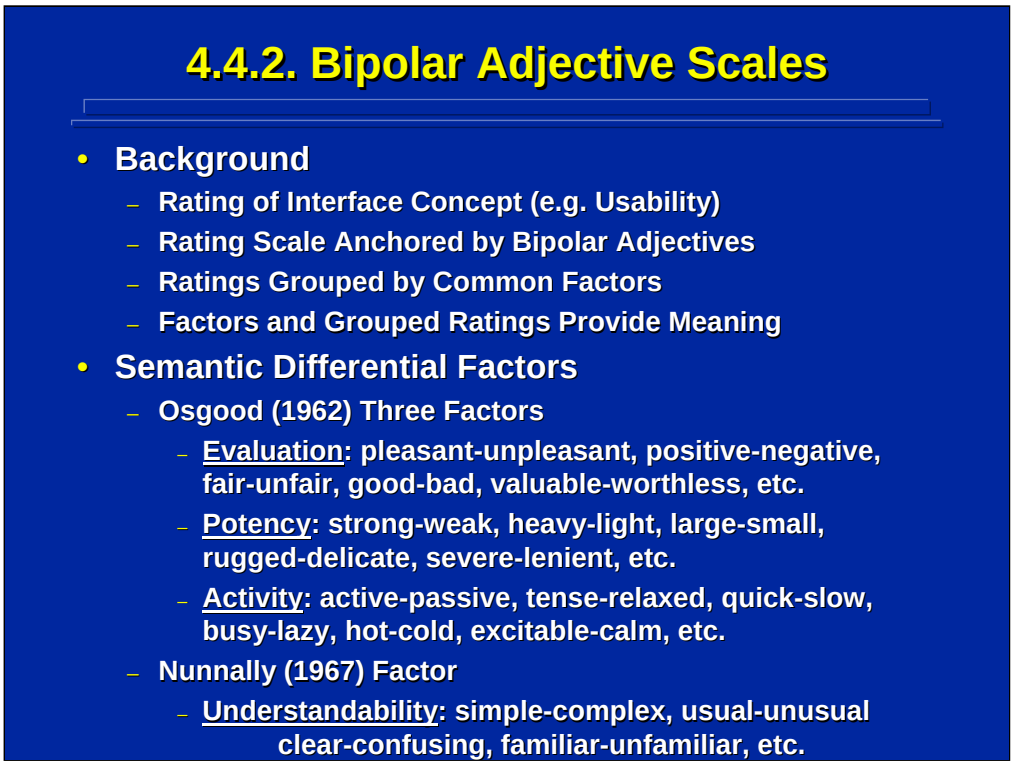
Graphic ratings are unbroken line or boxes with labeled divisions representing a characteristic, behavior, or dimension to be rated. The major demarcations on the scale are anchored with numbers and/or verbal labels. The subjects merely mark their answer directly on the scale. There are common formatting variations of graphical rating scales such as scale orientation, number of categories, order of scale, center point, and labels. Meister (1985) describes several characteristics of these parameters that need to be considered when designing and using ratings. The two major considerations of rating scales are the number of subdivisions and the labeling of scale anchors. Human factors researchers commonly use a horizontal scale with 3-to-9 categories designated by numbers along with verbal labels.




Likert (1932) developed a graphical rating scale that is probably the most frequently used rating in human factors research today. As shown on this slide, this scale consists of a statement followed by a horizontal scale with five categories. Below each number is a verbal label going from “Strongly Approve” to “Strongly Disapprove”. Many of the scales that one sees in research today are variations of the original Likert scale and are referenced as “Likert-type” scales.

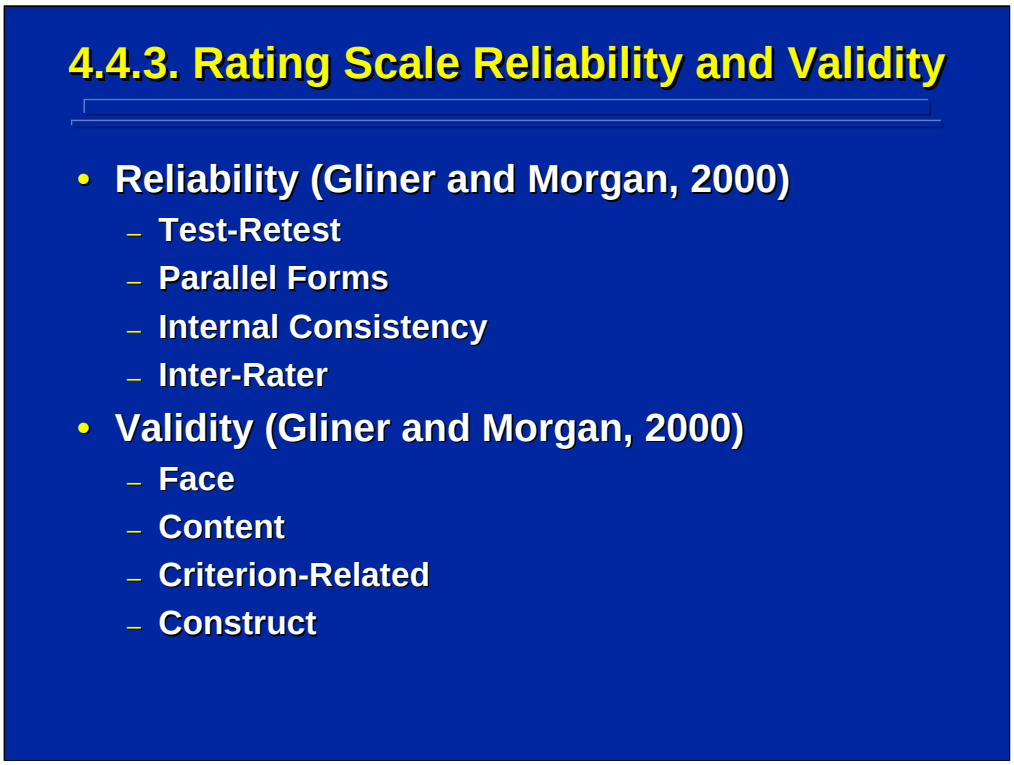
Some researchers argue that the Likert scale is set up with equal distances between the numbers, and therefore it represents interval scale data amenable to parametric analysis. Alternatively, one could argue that the difference between the adjectives may not be the same psychologically, and a nonparametric analysis should be used assuming ordinal data, at best.

Instructions and well planned procedures are critical in order to obtain consistent results using Likert-type ratings. The subject should be instructed to circle the number directly to avoid checking an answer somewhere between two numbers that the experimenter must then interpret as 4, 4.5, or 5, for example.




Another type of rating scale consisting of several bipolar adjectives (e.g. agree-disagree, good-bad) is also used as a way of obtaining supplemental data instead of the Likert-type scale that uses just one bipolar adjective, i.e. approve-disapprove.

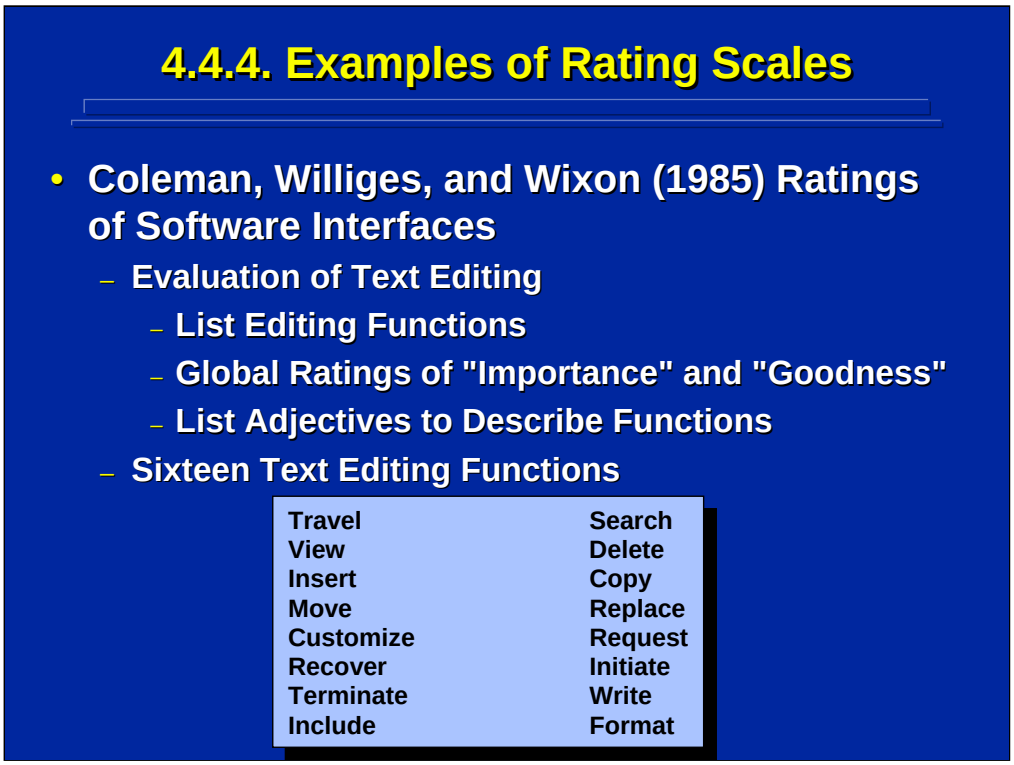
The resulting set of bipolar adjectives can be grouped into common factors that are meaningful. Osgood (1962) recommended three semantic differential factors that usually occur. They consist of evaluation, potency, and activity groupings of bipolar adjectives as shown on this slide. Alternatively, Nunnally (1967) suggested understandability as another factor grouping for bipolar adjectives. To build a rating scale of bipolar adjectives, one merely chooses a bipolar set from each dimension that is appropriate for the specific evaluation.




When a rating scale is used to collect supplemental data in only one study, various reliability and validity correlation coefficients are usually not calculated. See Gliner and Morgan (2000) for calculation details since they are beyond the scope of this reference material.

Reliability is the consistency of response. Measurements of reliability might include consistency when the same individual responds a second time (i.e. test-retest), consistency in parallel forms of the scale, internal consistency among items measuring the same concept (i.e., split-half reliability, Kuder-Richardson 20, and Cronbach's $\alpha$), and inter-rater consistency when two or more observers (i.e., experts) rate subjects in the experiment. Often just the correlation among multiple raters is calculated to determine consistency among raters of supplemental data.

Validity determines if the rating scale really measures what it is supposed to measure. Conceptually, various dimensions of measurement validity are considered that can include determining if appearance of material has relevance to the rater (i.e. face validity), determining if the actual content of the rating scale is relevant to the concept being evaluated, validating the scale against an external criterion, and determining how well a rating scale actually measures an underlying construct such as usability.



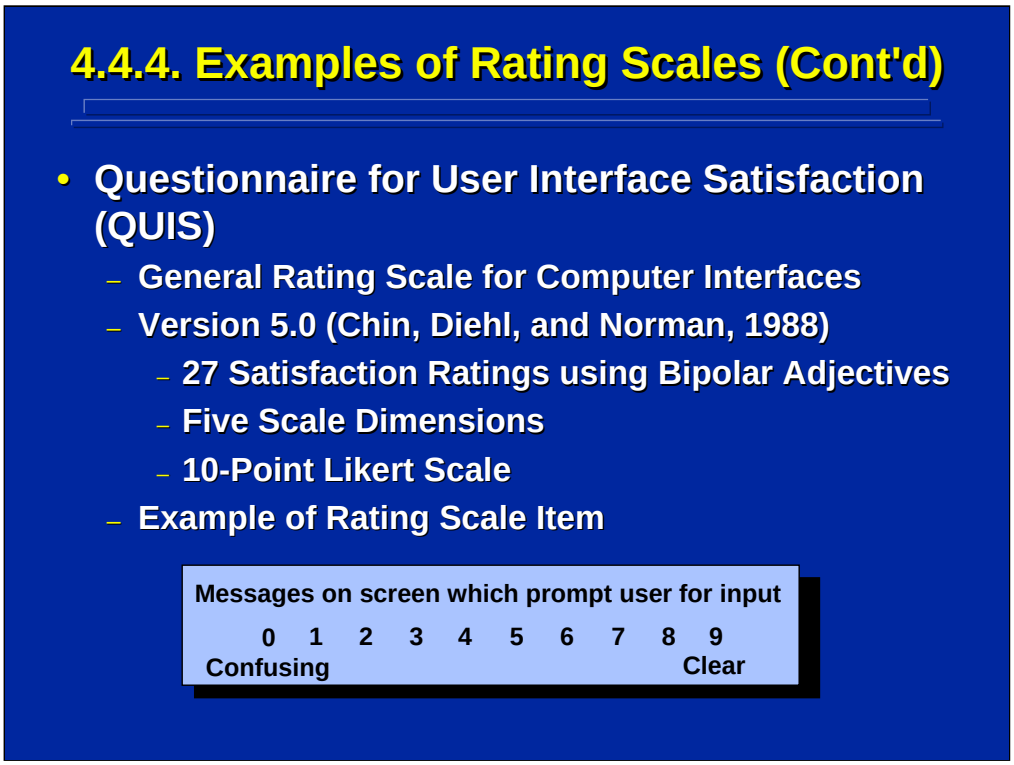

Two examples are provided for using rating scales as a means of collecting subjective data in human factors research. Both examples deal with problems related to human-computer software interface design. The first example shows the bipolar adjective scales developed by Coleman, Williges, and Wixon (1985) to measure the “importance” and “goodness” of various text editing functions. They evaluated the 16 editing functions shown on this slide. In addition to evaluating the effects of these functions on text editing performance in terms of speed and errors, they also looked at the supplemental data on subjects’ evaluations of goodness and importance of these functions. In the process, they attempted to develop a general rating scale to evaluate text editors.

## 4.4.4. Examples of Rating Scales (Cont'd)

- Coleman, et al. (1985) Rating Scale Development
  - Descriptive Adjectives
    - 86 Initial Adjectives
    - 28 Principal Components
    - Rated on 7-Point Scale of "Importance"
    - 17 Highest Rated “Importance” Adjectives
  - Rating Scale Characteristics
    - Seven-Point Likert-Type Scale
    - Anchored by Bipolar Adjectives
- Evaluation of Editing Functions
  - All 17 Bipolar Adjective Rated with "Goodness"
  - Provides More Detailed Description of "Goodness"

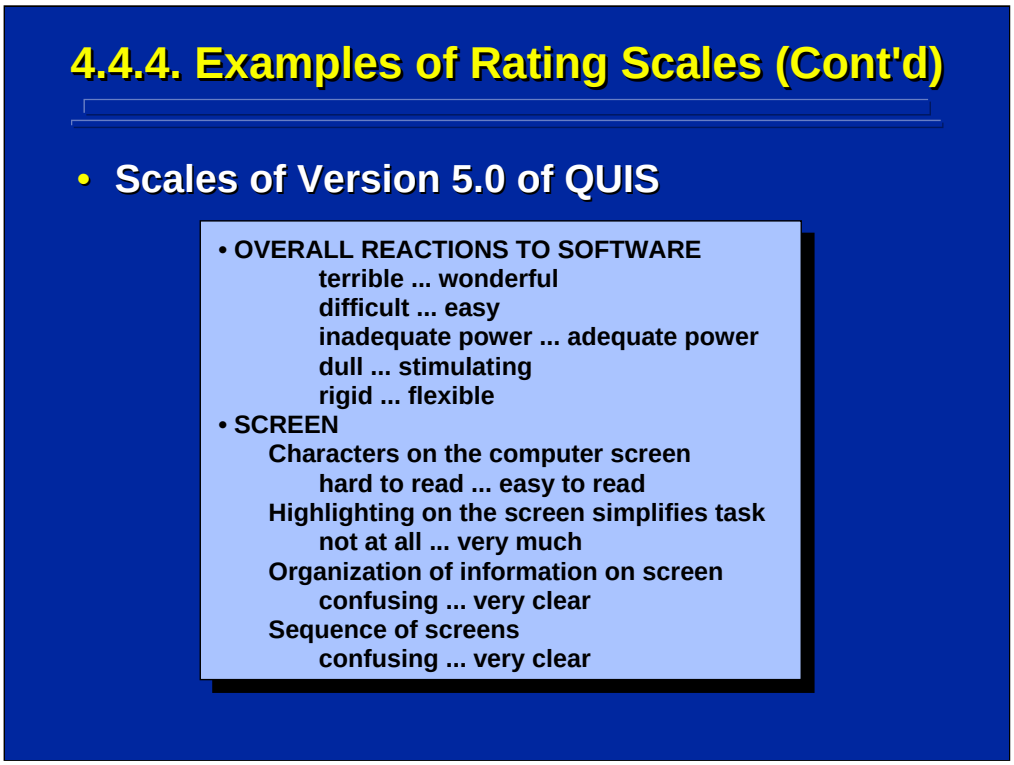
Coleman et al. (1985) used 86 bipolar adjectives that could group into 28 different groups. Then they developed a 7-point, Likert-type rating scale of importance and isolated the top rated 17 “importance” bipolar adjectives. Each of the resulting 17 scale items they selected appeared as a 7-point Likert-type scale that was anchored by bipolar adjectives. All 17 bipolar adjectives were also used to provide “goodness” ratings.

This slide summarizes the 17 bipolar adjectives used in the 7-point, Likert-type scales by Coleman et al. (1985) to evaluate importance and goodness of various text editing functions. Note that the rank orders of the ratings using the 17 bipolar adjectives depend upon whether one is assessing importance or goodness. Consequently, these rank orders can be used to help interpret what importance and goodness ratings of text editing functions mean. For example, “dependable”, “useful”, and “fast” are the three highest rated adjectives in evaluating importance; whereas “pleasing”, “friendly”, and “complete” are the three highest rated adjectives for evaluating goodness.

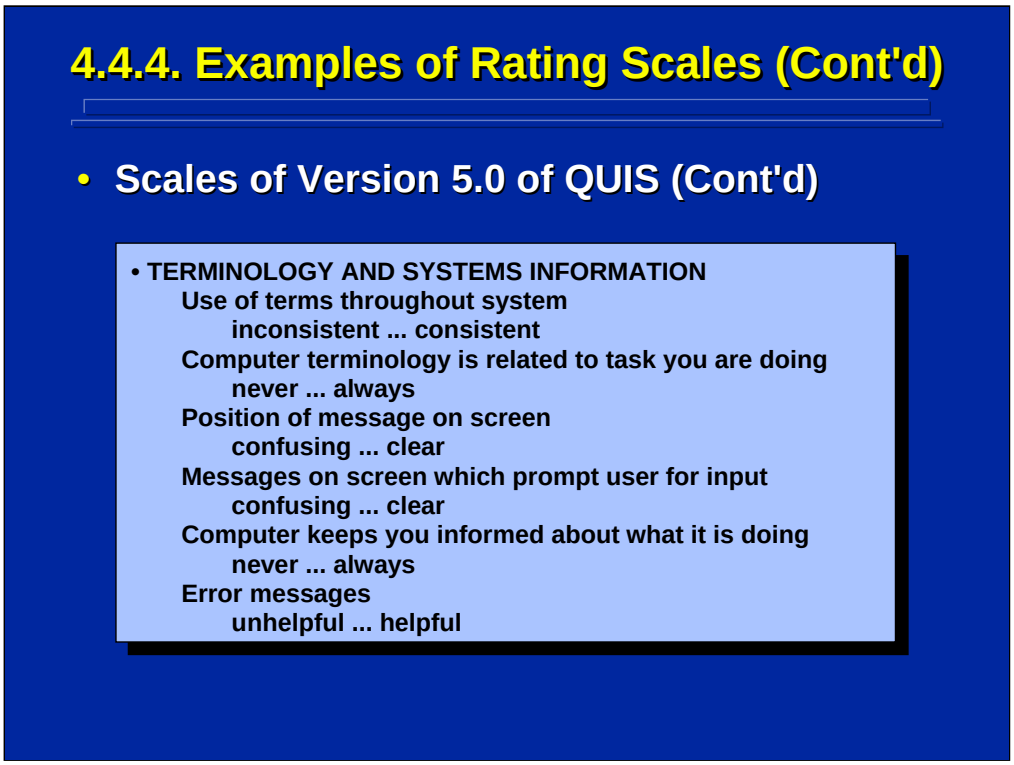



The second example of a rating scale is the Questionnaire for User Interface Satisfaction (QUIS) by Chin, Diehl, and Norman (1988). This scale was developed as a general satisfaction rating scale for computer interfaces. Researchers potentially can make comparisons across many studies and families of computer interfaces using a standard metric of user satisfaction using QUIS.

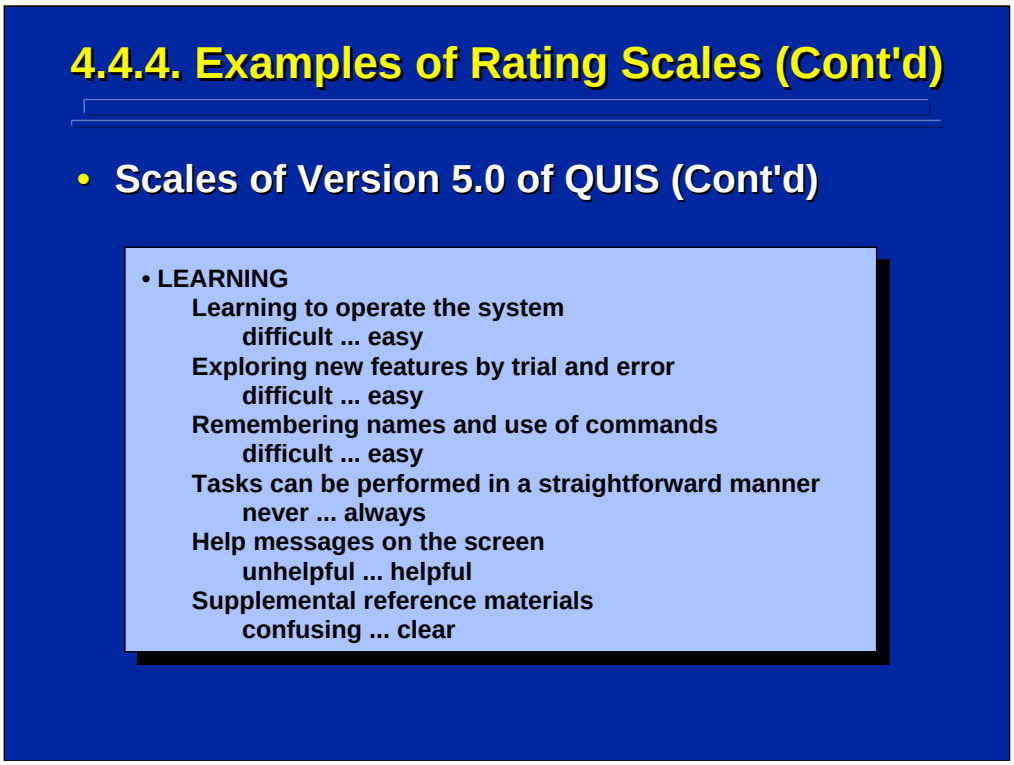
Chin, et al. (1988) described Version 5.0 of QUIS in a proceedings paper at a technical conference; however, subsequent versions are only commercially available. They used 27 satisfaction ratings grouped into 5 interface dimensions. Each of the 27 ratings is made on a 10-point Likert-type scale. The bottom of this slide shows an example of one of these 27 scale items.



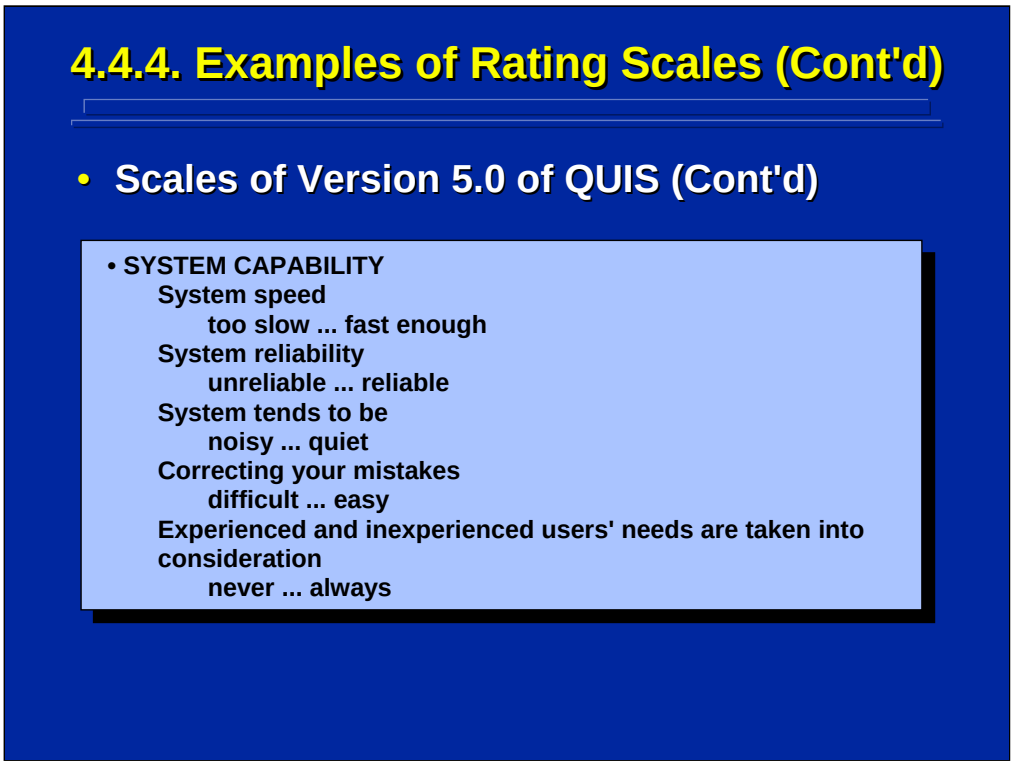

The next four slides show the five dimensions of Version 5.0 of QUIS. This slide lists the first two dimensions of the QUIS. The first dimension is the overall reaction to the software and consists of 5 bipolar adjective ratings. The second dimension is the screen evaluation based on 4 ratings.



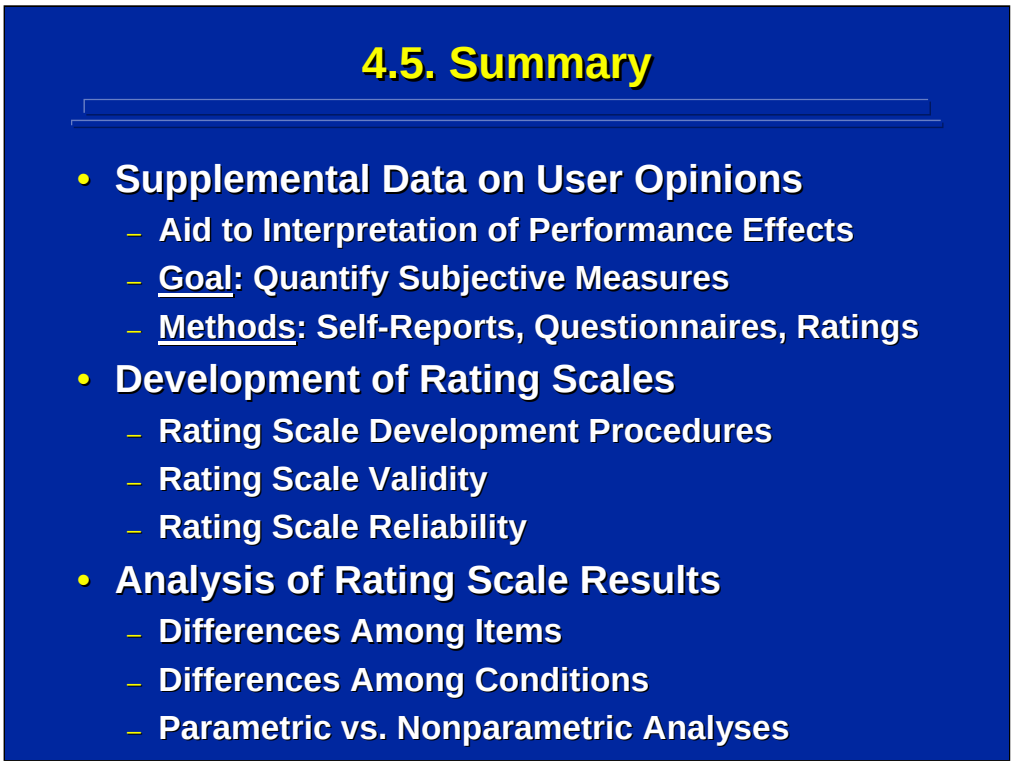

This slide summarizes the 6 ratings used to evaluate the terminology and the systems information of Version 5.0 of QUIS.




The fourth dimension of Version 5.0 of QUIS relates to learning the interface and consists of 6 ratings shown on this slide.



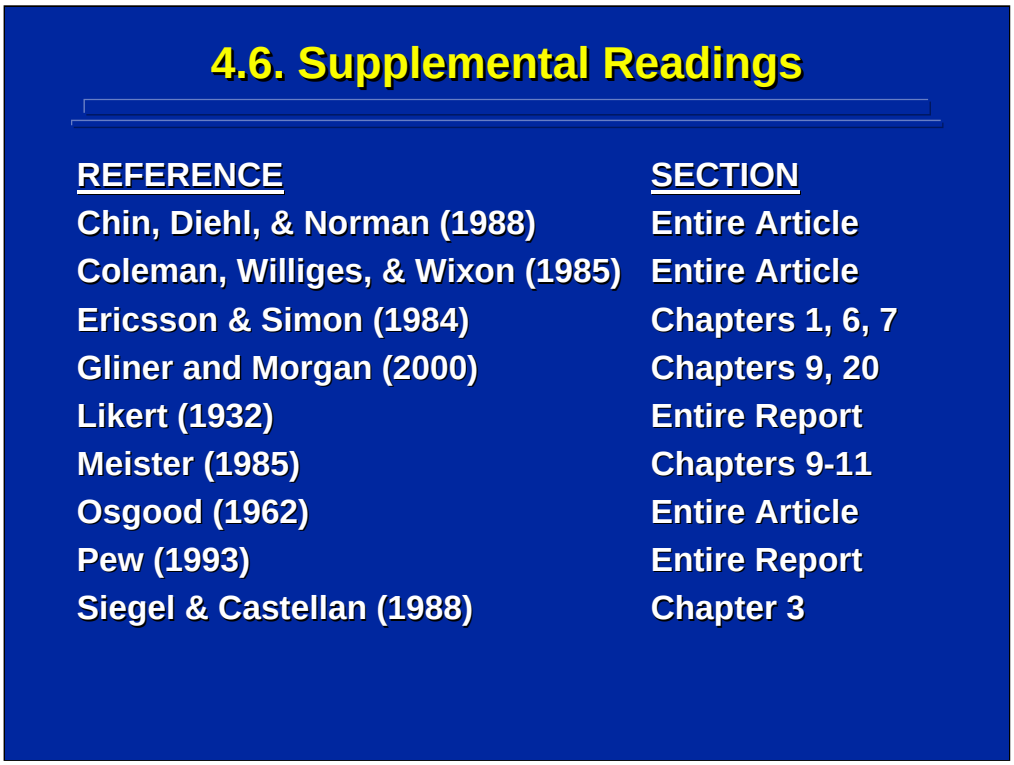

The fifth dimension of Version 5.0 of QUIS deals with the system capability and is evaluated by the 5 ratings shown on this slide.



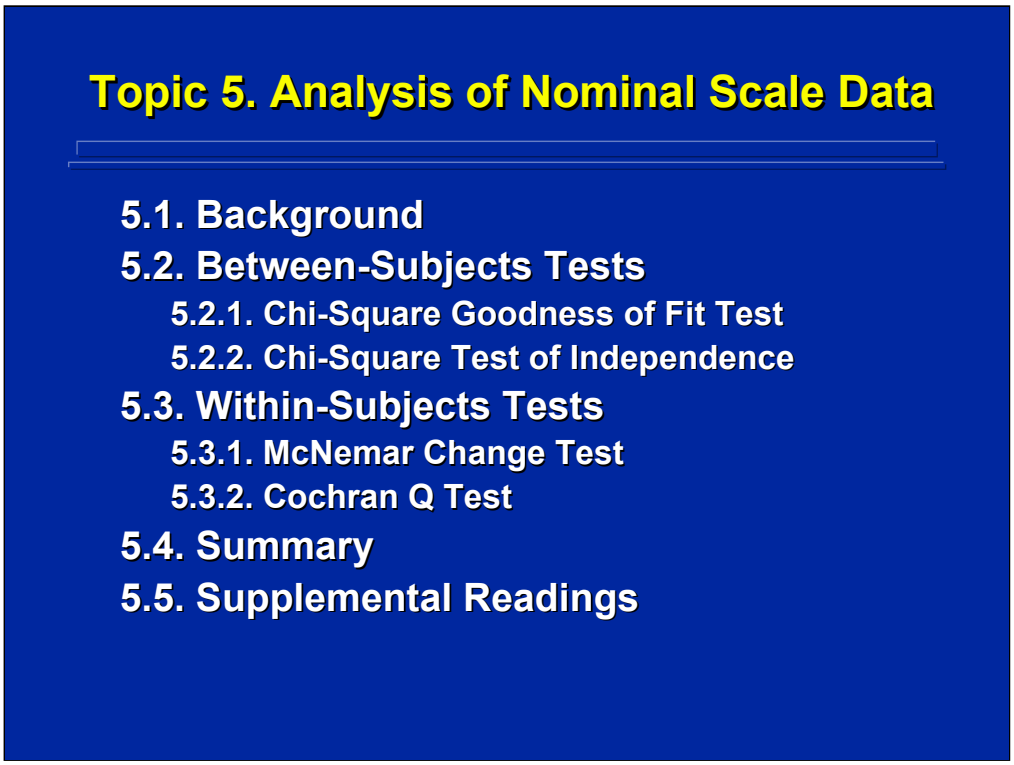

By way of summary, remember that the goal of supplemental data collection is to aid interpretation of the overall performance effects evaluated through experimental design. To facilitate this process, one should try to quantify the subjective measures as much as possible.

Three often used methods for collecting supplemental data in human factors and ergonomics research are self-reports, questionnaires, and rating scales. Of these three the one most amenable to subsequent quantitative analysis is the rating scale. If the same rating scale is going to be used across a series of research efforts, then the researcher should consider using a systematic development process to determine both the validity and reliability of the rating scale.

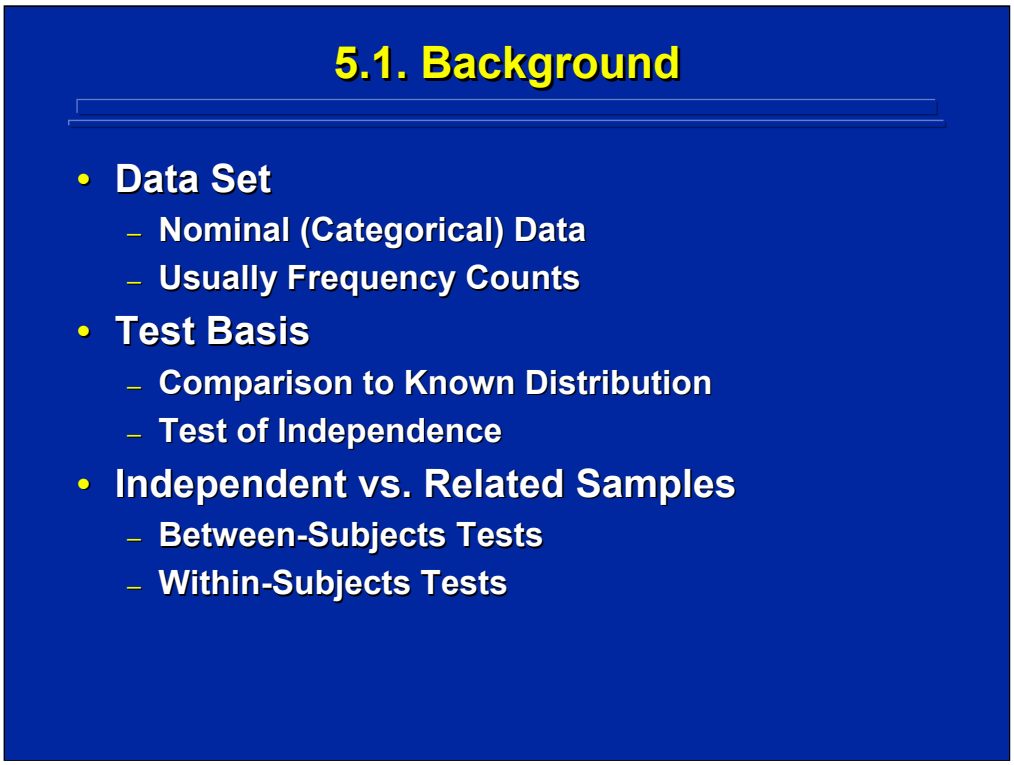
The analysis of supplemental data includes both the analysis of responses to different items or ratings and the analysis of summary ratings across dimensions for different treatment conditions in an experiment. Since most supplemental data has only nominal or ordinal properties, at best, the analyses use nonparametric procedures.



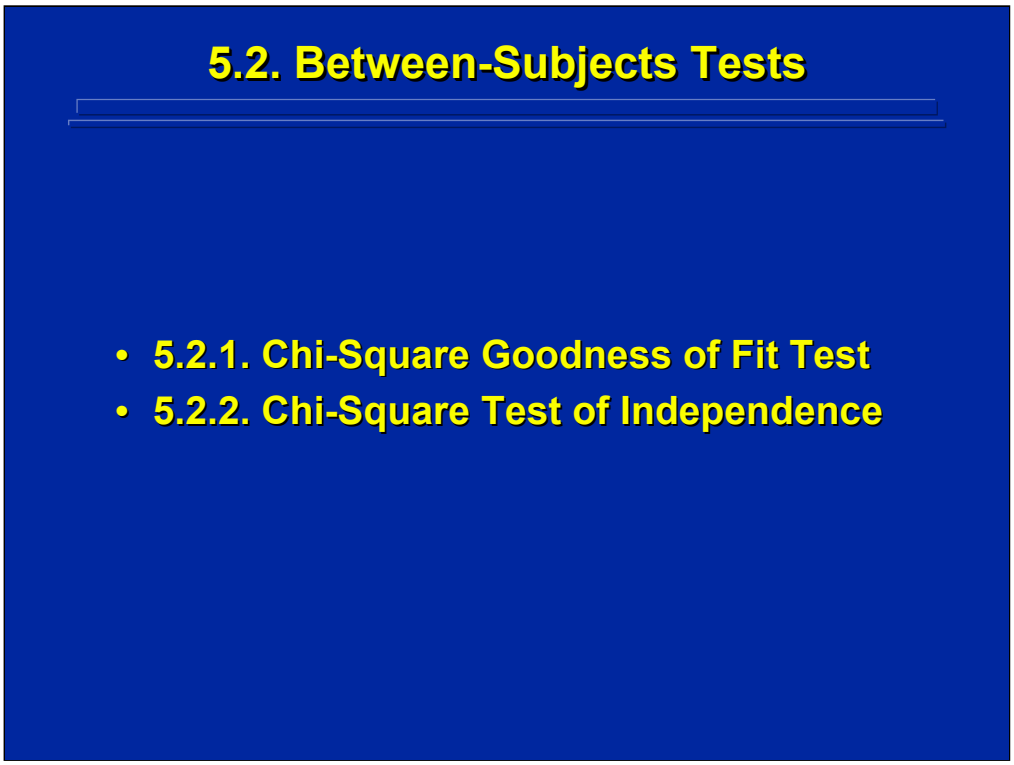

Meister (1985) discusses general issues related to rating scales, and Pew (1993) discusses general issues in the collection of subjective data designed specifically for human factors and ergonomics research. The chapter suggested in Siegel and Castellan (1988) provides a description of measurement scales and an introduction to nonparametric analysis. Gliner and Morgan (2000) provide details on scales of measurement as well as reliability and validity of measurements for rating scales. The other references give the details of specific techniques that are reviewed in this topic for collecting supplemental data.



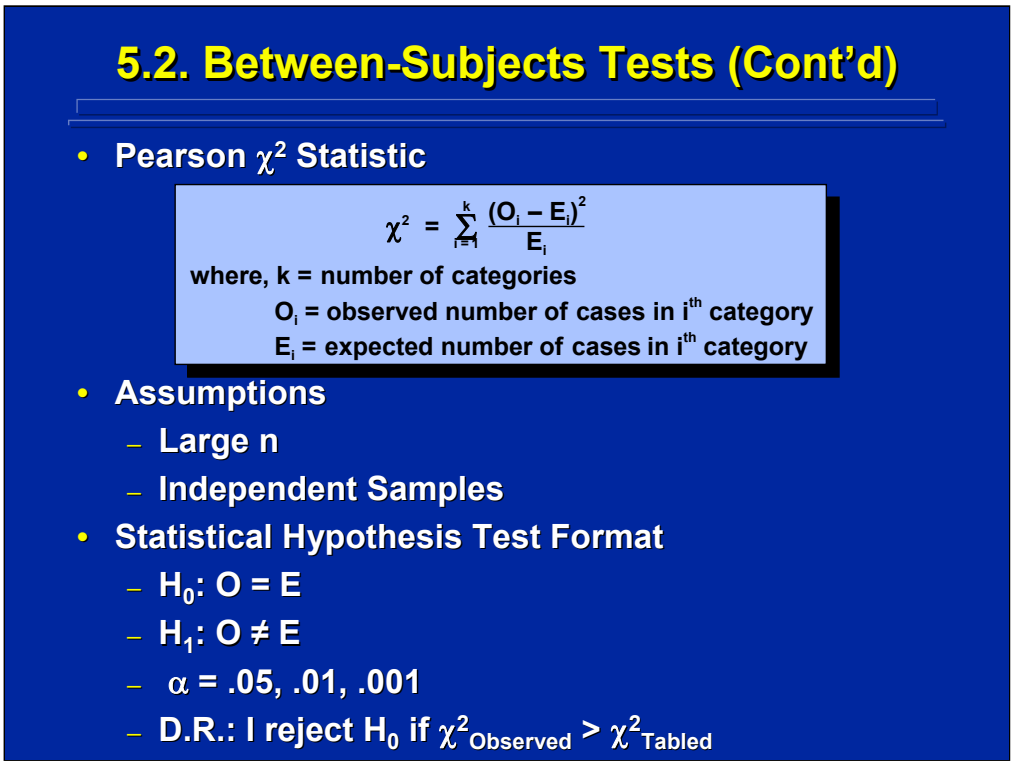

This topic deals with an overview of major supplemental data analysis alternatives that can be used with nominal scale data. Only a sample of nonparametric analyses covering the most common techniques used in human factors and ergonomics research is presented in this topic. Siegel and Castellan (1988) provide a detailed discussion of all of these techniques, and their formulae and notation are used throughout this topic for easy reference. The nonparametric techniques in this reference are organized around between-subjects and within-subjects techniques to facilitate an easy choice of nonparametric analysis procedures for nominal data.




If supplemental data are only nominal scale, one has categorical data that is usually in the form of frequency counts within a category. The test of statistical significance is either a comparison of the supplemental data to some known distribution or a test of independence among the different categories. The choice of test alternative depends on whether the data are between-subjects or within-subjects observations.




This subsection summarizes two between-subjects nonparametric tests for categorical data. Both the goodness of fit test and the test of independence among two or more categories use the chi-square sampling distribution. So, both are referred to as chi-square tests.




The Pearson Chi-Square statistic for discrete categorical data is used to approximate the continuous chi-square sampling distribution assuming large and independent samples (Hays and Winkler, 1971, p. 784; Hays, 1994, p. 862). The formula for the Pearson statistic is shown on the slide. It is simply the sum of the observed value (O) minus the expected value (E) squared divided by E. The value of O is the frequency count in each category of the actual supplemental data, but the E value depends upon the type of hypothesis test the researcher is conducting. Procedures for calculating E for both a goodness of fit test and a test for independence will be described in this reference material.

The standard hypothesis-testing format can be used for either test. This format is shown at the bottom of this slide in terms of O and E using the Pearson Chi-Square statistic.

## 5.2.1. Chi-Square Goodness of Fit Test

- **Background**
  - Expected Frequency (E) Defined by Known Distribution
  - Observed Frequency (O) Defined by Sample Data
- **Required Sample Size**
  - Rule of Thumb: E ≥ 5 for Each Category
  - Combine Categories When df > 1
  - Yate's Correction When df = 1

$$\chi^2_{Observed} = \sum_{i=1}^{k} \frac{(|O - E| - .5)^2}{E}$$

  - Binomial Test

A goodness of fit test compares the observed value, O, resulting from the supplemental data collection to known population values of categories, E, in order to calculate the Pearson Chi-Square statistic. A rule of thumb is that the E in each of the categories should be greater than or equal to 5 in order for the Pearson chi-square to approximate the chi-square distribution adequately. If E is less than 5 one can combine categories, if meaningful, to provide an E equal to or greater than 5.

If E is still not equal to or greater than 5 when the goodness of fit test reduces to only 2 categories, (i.e., one degree of freedom), Hays and Winkler (1971, p. 788) suggest that one can use a Yate's correction as shown on this slide when the degrees of freedom are equal to 1. This correction, however, could be quite conservative (Delucchi, 1993, p. 304). Alternatively, Siegel and Castellan (1988. p. 50) recommend using a binomial test instead of a Pearson Chi-Square when E is less than 5.

## 5.2.1. Chi-Square Goodness of Fit Test (Cont'd)

- **Example Problem: The relative frequency of the age of automobile drivers in the U.S. is known. A sample of 50 drivers is chosen, and demographic data on age is recorded. Does the age of this sample differ from the distribution of the U.S. population of drivers ($p < 0.01$)?**

| Age of Driver | U.S. Population | Expected (E) | Observed (O) |
|---|---|---|---|
| 18-25 | 0.19 | 9.5 | 10 |
| 26-35 | 0.11 | 5.5 | 3 |
| 36-45 | 0.15 | 7.5 | 6 |
| 46-55 | 0.27 | 13.5 | 25 |
| 56-65 | 0.16 | 8 | 5 |
| >65 | 0.12 | 6 | 1 |
| | 1.00 | 50 | 50 |

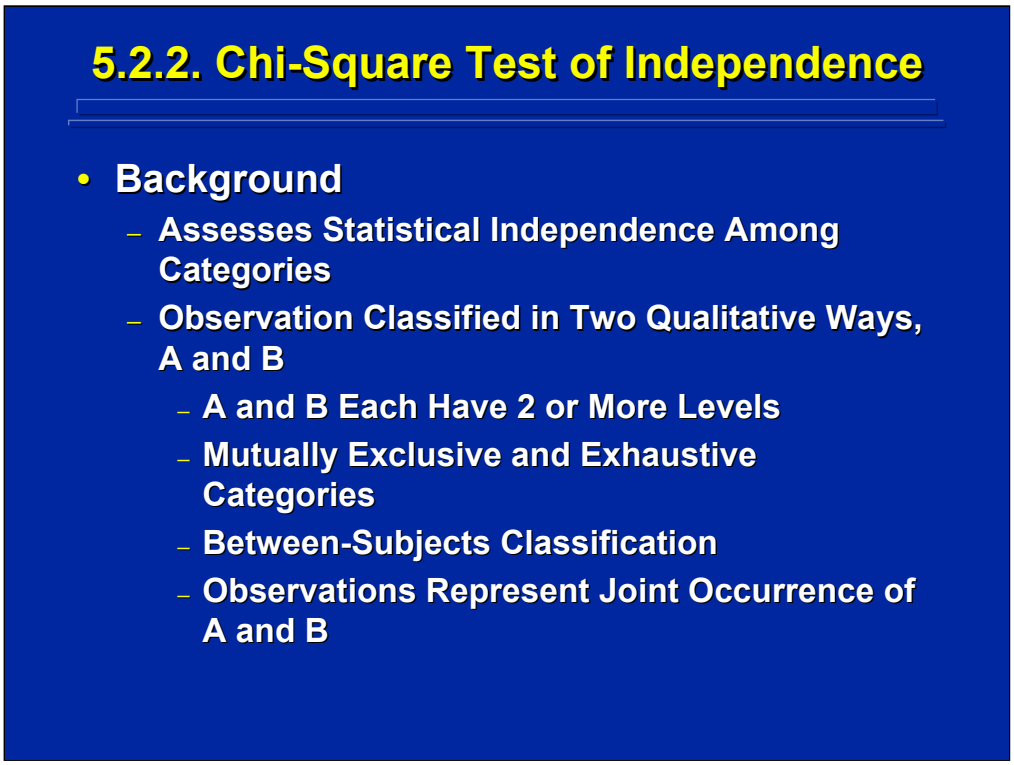
$$\chi^2_{Observed} = \frac{(10-9.5)^2}{9.5} + \frac{(3-5.5)^2}{5.5} + \frac{(6-7.5)^2}{7.5} + \frac{(25-13.5)^2}{13.5} + \frac{(5-8)^2}{8} + \frac{(1-6)^2}{6} = 16.55^*$$

$$\chi^2_{Tabled} = (6-1) = 5\ df = 15.09$$

(Click in this red rectangle to see SAS calculations for this example.)

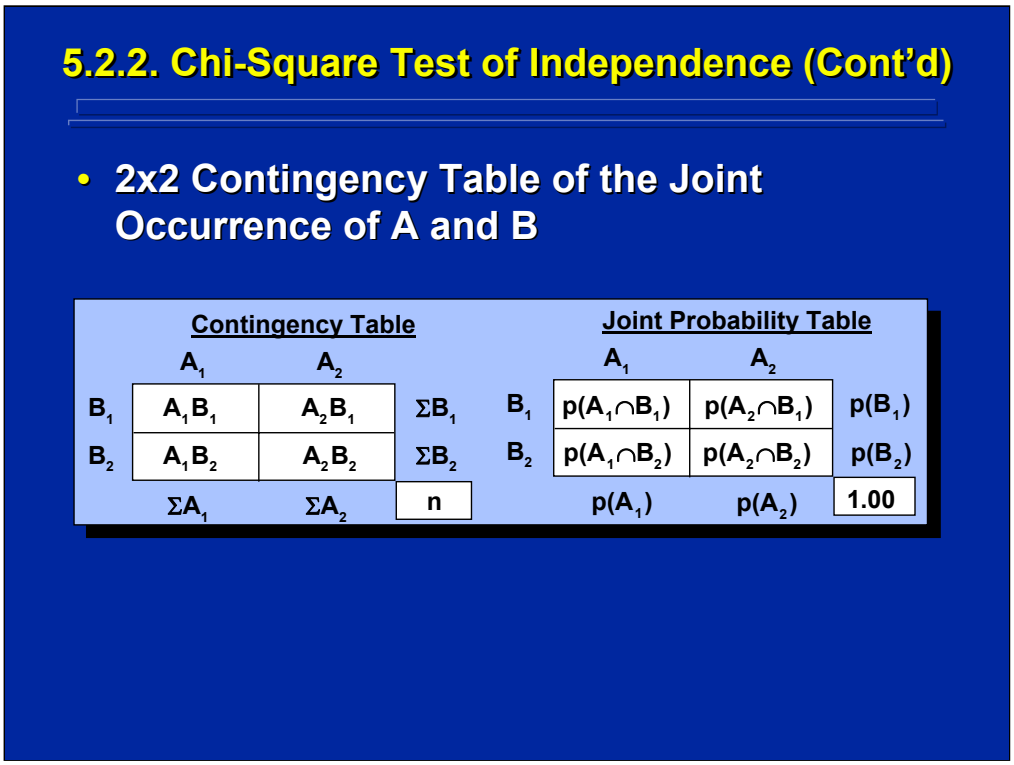
A hypothetical example of a chi-squared goodness of fit test is shown on this slide. The U.S. driving population is converted to the expected frequency, E, based on a sample size of 50. The values of E are tested against the observed frequencies in each driver age category, O, using a chi-squared test of significance. Since all values of E in this example are greater than five, one can probably use the Pearson Chi-Square statistic without correction.

The resulting chi-square observed is compared to a tabled value with 5 degrees of freedom (i.e, 6 categories) resulting in a significant difference ($p < 0.01$). Therefore, age distribution of the sample of 50 drivers used in this study is significantly different from the age distribution of drivers in the U.S. population.

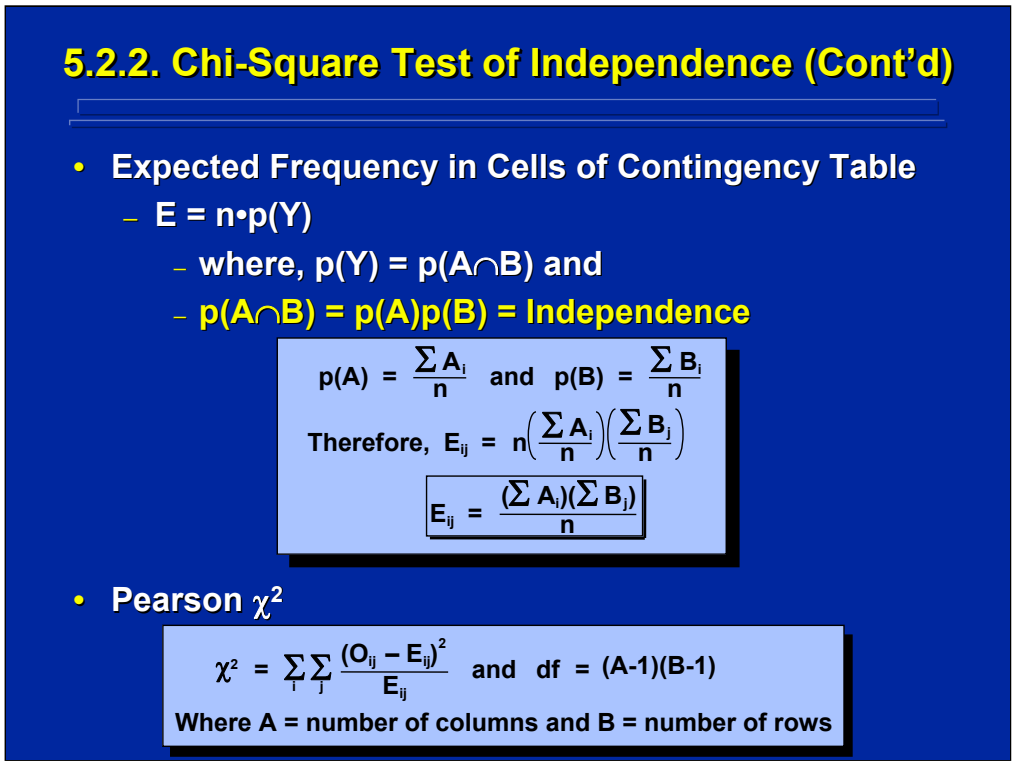



A chi-squared test for independence is a useful significance test for comparing frequencies classified in two qualitative ways, A and B. Each observation can be classified in terms of A and B, where both A and B have two or more levels. The classifications are mutually exclusive and exhaustive categories that result in contingency tables of the frequency of occurrence of between-subjects observations.

The resulting significance test is based on the joint occurrence of the levels A and B in the cells of the resulting contingency table of the various AB combinations. This test also uses the Pearson Chi-Square statistic and determines the expected frequency, E, based on the assumption of statistical independence between A and B.




This slide shows a 2x2 contingency table beside its corresponding joint probability table. The sum of the cells of the column in the contingency column is the sum of $A_1$ or $A_2$. The same goes for the rows of the table in terms of $B_1$ or $B_2$. One can estimate the probability of these frequency counts in each cell, which is the four joint probabilities of A and B shown on this slide.




Under the assumption of independence, these joint probabilities of A and B equal the p(A) times p(B). The formula shown in the center of this slide shows that the expected frequency ($E_{ij}$) of every cell in a contingency table can be estimated based on the assumption of independence. The resulting values of E are then used in the Pearson Chi-Square formula shown on the bottom of this slide for calculating chi-square observed in a test of independence.

## 5.2.2. Chi-Square Test of Independence (Cont'd)

- **Alternate Form of Pearson $\chi 2$ for 2x2 Contingency Table**

**Contingency Table**

| | $A_1$ | $A_2$ | |
|---|---|---|---|
| $B_1$ | a | b | a + b |
| $B_2$ | c | d | c + d |
| | a + c | b + d | n |

$$\chi^2_{Observed} = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

**Yate's Correction When Any $E_{ij} < 5$**

$$\chi^2_{Observed} = \frac{n(|ad - bc| - n/2)^2}{(a + b)(c + d)(a + c)(b + d)}$$

Siegel and Castellan (1988, pp.116-117) provide formulae for calculating the Pearson Chi-Square directly from observed frequencies without having to calculate the joint probability in a 2x2 contingency table. This slide shows the alternate formula for the general Pearson Chi-Square statistic as well as the Yates correction when any cell frequency in a contingency table is equal to or less than 5.

## 5.2.2. Chi-Square Test of Independence (Cont'd)

- **Example Problem: Every user in a random sample of 80 users classified themselves as either high (Hi) or low (Lo) in computer experience. All users practiced using an experimental text editor for 10 hours and were then asked to state whether they were satisfied (Yes) or not satisfied (No) with the text editor. Is their satisfaction evaluation independent of their computer experience ($p < 0.05$)?**

**Observed Frequency, $O_{ij}$**

| | Yes | No | |
|---|---|---|---|
| Hi | 24 | 11 | 35 |
| Lo | 16 | 29 | 45 |
| | 40 | 40 | 80 |

**Expected Frequency, $E_{ij}$**

| | Yes | No |
|---|---|---|
| Hi | 17.5 | 17.5 |
| Lo | 22.5 | 22.5 |

Where, $E_{11} = \frac{(40)(35)}{80} = 17.5$

(Click in this red rectangle to see SAS calculations for this example.)

This is an example of the chi-squared test for independence. Each of the 80 subjects who used the experimental text editor rated their computer experience as either high or low. The researcher is interested in determining whether or not computer experience influences user satisfaction with the experimental editor being evaluated. This is a between-subjects test since each subject was classified into only one level of experience.

The observed data shown in the left-hand table on the slide are frequency counts. Consequently, the appropriate test for the hypothesis in question is a chi-square test of independence. The expected values, E, based on independence are shown in the right-hand table on this slide.

## 5.2.2. Chi-Square Test of Independence (Cont'd)

- **Statistical Hypothesis Test**
  - $H_0$: O = E
  - $H_1$: O ≠ E
  - $\alpha = .05$
  - D.R.: I reject $H_O$ if $\chi^2_{Observed} > \chi^2_{Tabled}$

$$\chi^2_{Obs} = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(24 - 17.5)^2}{17.5} + \ldots + \frac{(29 - 22.5)^2}{22.5} = \boxed{8.58}$$

$$\chi^2_{Tab} = [df = (2 - 1)(2 - 1) = 1] = \boxed{3.84}$$

- **Summary of Chi-Square Contingency Tables**
  - Calculate $E_{ij}$ Based on Independence
  - Can Be Extended Beyond 2x2 Contingency Tables

(Click in this red rectangle to see SAS calculations for this example.)

The observed and expected frequencies are used to calculate the observed Pearson Chi-Square statistic shown on this slide. The standard format can be used to summarize the hypothesis test. Since the chi-square observed is larger than the chi-squared tabled value, one can conclude that there is a significant difference which means that user computer experience and text editor preference are not independent. So, satisfaction with the experimental text editor depends on the user's computer experience.

To calculate any test of independence, one uses the Pearson Chi-Square and calculates the expected frequency based on the assumption of independence. The 2x2 contingency table shown in this example can be generalized to larger contingency tables.

## 5.2.2. Chi-Square Test of Independence (Cont'd)

- RxC Contingency Tables
  - Sample Size: n>20 and $E_{ij}>5$
  - Procedure
    - Calculate $E_{ij}$ by Appropriate Marginal Totals, $R_i$ and $C_j$ where $E_{ij} = R_i C_j / n$
    - Observed Pearson $\chi^2$

$$\chi^2_{Observed} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \text{and} \quad df = (r-1)(c-1)$$

  - Isolate Significant Effects
    - Partition RxC Contingency Table into Series of 2x2 Tables Each Having 1 df
    - Additive 2x2 Tables
    - Partition by Meaningfulness

Expanded 2x2 contingency tables are stated in terms of RxC tables where R is the number of categories represented in the rows of the contingency table and C is the number of categories represented in the columns of the contingency table. In these larger contingency tables, one can use the Pearson Chi-Square statistic shown on this slide and the chi-square sampling distribution if the sample size is greater than 20 and the expected values are greater than 5.

Any significance found in the expanded RxC contingency table merely states that some of the joint frequencies are not independent, but it does not specify exactly where this lack of independence occurs. To isolate the significant effects one can partition the overall RxC contingency table into a series of smaller 2x2 contingency tables, each having 1 degree of freedom, for subsequent analysis. The choice of meaningful 2x2 partitions then helps isolate significant effects.

## 5.2.2. Chi-Square Test of Independence (Cont'd)

- Example of RxC Contingency Table
  - Example Problem: Previous example with 80 subjects divided into "Hi", "Med", and "Lo" computer experience
  - Overall Test of Significance

**3x2 Contingency Table**

| | | |
|---|---|---|
| $O_{11}$ ($E_{11}$) | $O_{12}$ ($E_{12}$) | $R_1$ |
| $O_{21}$ ($E_{21}$) | $O_{22}$ ($E_{22}$) | $R_2$ |
| $O_{31}$ ($E_{31}$) | $O_{32}$ ($E_{32}$) | $R_3$ |
| $C_1$ | $C_2$ | n |

**Results**

| | Yes | No | |
|---|---|---|---|
| Hi | 24 (17) | 10 (17) | 36 |
| Med | 8 (7.5) | 7 (7.5) | 13 |
| Lo | 8 (15.5) | 23 (15.5) | 31 |
| | 40 | 40 | 80 |

where $E_{11}=R_1C_1/n$ ... $E_{ij}=R_iC_j/n$

$\chi^2_{Obs} = (24-17)^2/17 + ... + (23-15.5)^2/15.5 = 13.09$

$\chi^2_{Tab} = (3-1)(2-1) = 2$ df $= 5.99$ $(p < 0.05)$

(Click in this red rectangle to see SAS calculations for this example.)

This example expands computer experience into 3 levels, High, Medium, and Low, as compared to the original example that used only 2 levels of computer experience, High and Low. In this example, each of the 80 subjects who used the experimental text editor rated their computer experience as high, medium, or low instead of just high or low experience in the previous example.

Consequently, one now has a 3x2 contingency table of user computer experience and text editor satisfaction. The calculations shown on this slide demonstrate an overall significance in the 3x2 contingency table. Again, one concludes that satisfaction with the experimental text editor depends upon computer experience. But, the differences between the three levels of computer experience on text editor satisfaction cannot be isolated by this overall test of significance because more than two levels of experience were evaluated. So, one must conduct subsequent 2x2 contingency table tests to isolate the locus of significance.

## 5.2.2. Chi-Square Test of Independence (Cont'd)

- **Isolating Significant Effects**
  - **Partition into (r-1)(c-1), 1 df, 2x2 Contingency Tables**
  - **Two Additive 2x2 Partitions of The Example 3x2 Contingency Table, [1] and [2]**

[1]

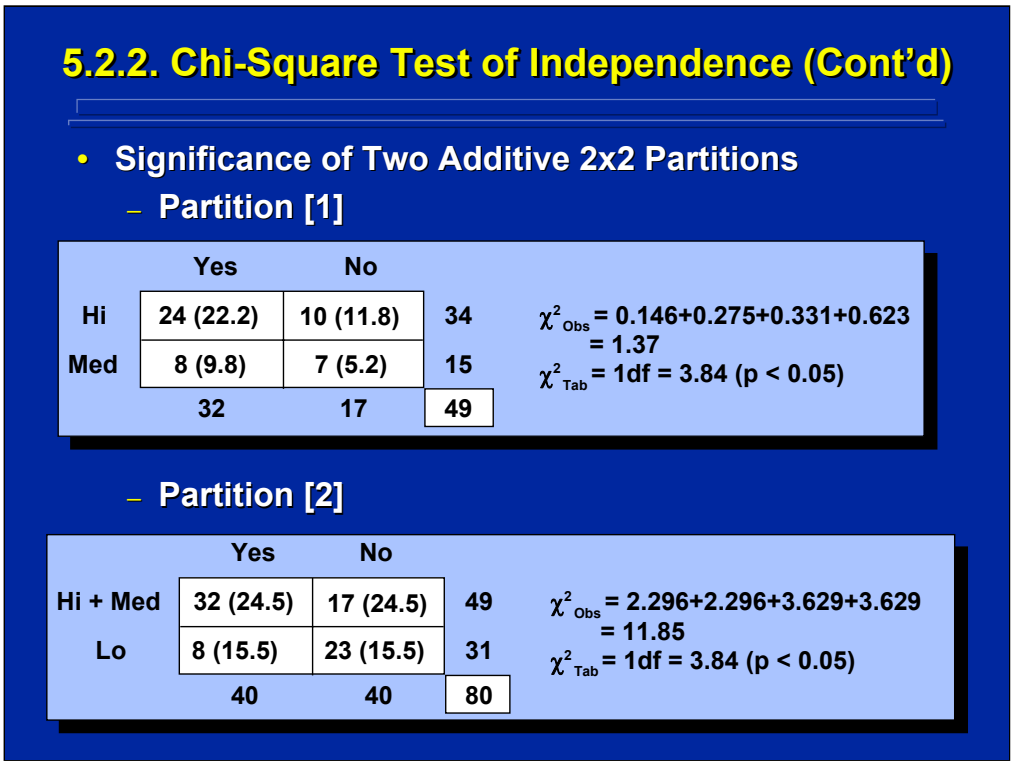
| | | |
|---|---|---|
| $O_{11}$ ($E_{11}$) | $O_{12}$ ($E_{12}$) | $R_1$ |
| $O_{21}$ ($E_{21}$) | $O_{22}$ ($E_{22}$) | $R_2$ |
| $C_1$ | $C_2$ | n |

[2]

| | | |
|---|---|---|
| $O_{11} + O_{21}$ ($E_{11} + E_{21}$) | $O_{12} + O_{22}$ ($E_{12} + E_{22}$) | $R_1 + R_2$ |
| $O_{31}$ ($E_{31}$) | $O_{32}$ ($E_{32}$) | $R_3$ |
| $C_1$ | $C_2$ | n |

(Click in this red rectangle to see SAS calculations for this example.)

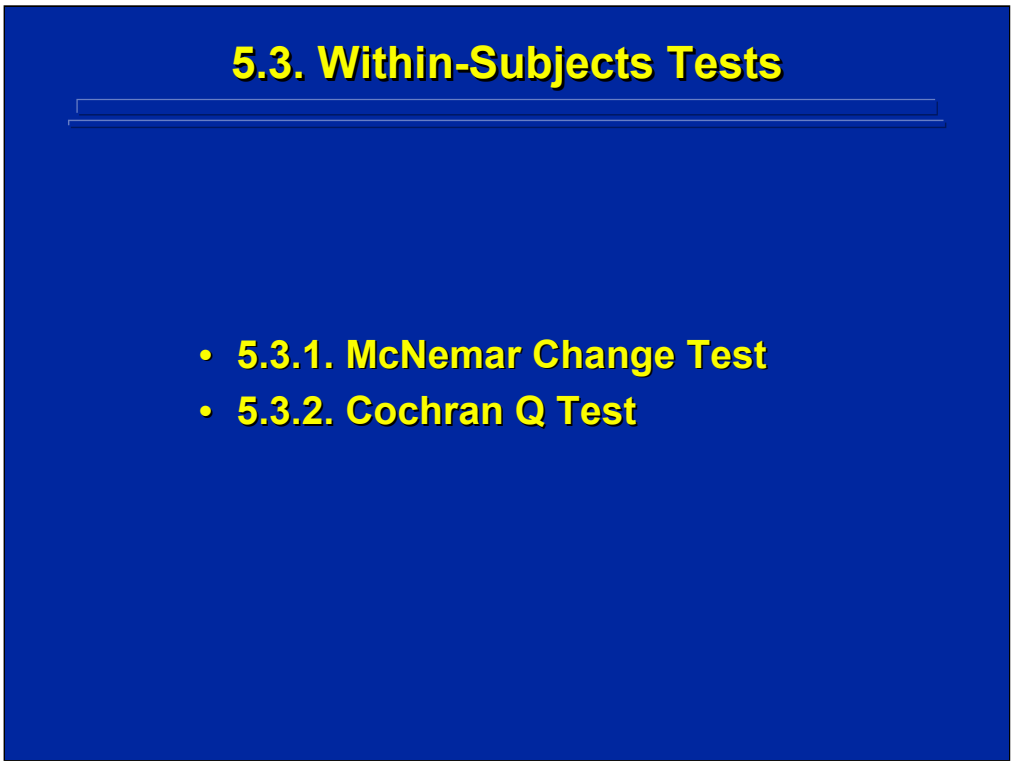
This slide shows a breakdown of the original 3x2 contingency table into two subsequent 2x2 contingency table tests. The first compares just High and Medium computer experience, and the second compares the additive effect of High and Medium experience to Low computer experience.



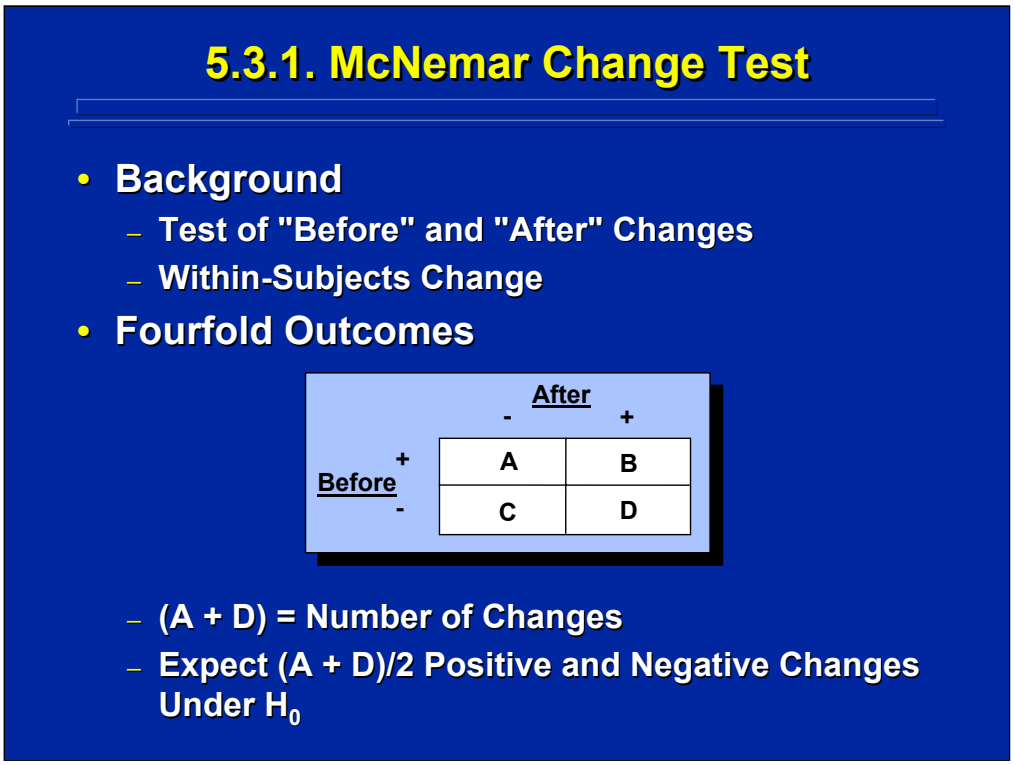

(Click in this red rectangle to see SAS calculations for this example.)

The calculations shown on this slide show no significant difference between High and Medium computer experience on text editor satisfaction, but the second 2x2 partition shows a significant difference. The second 2x2 contingency table combines High and Medium computer experience and compares this combination to Low computer experience. Consequently, the locus of difference in text editor satisfaction is between users with Low computer experience and users with more computer experience (i.e., Medium or High computer experience).

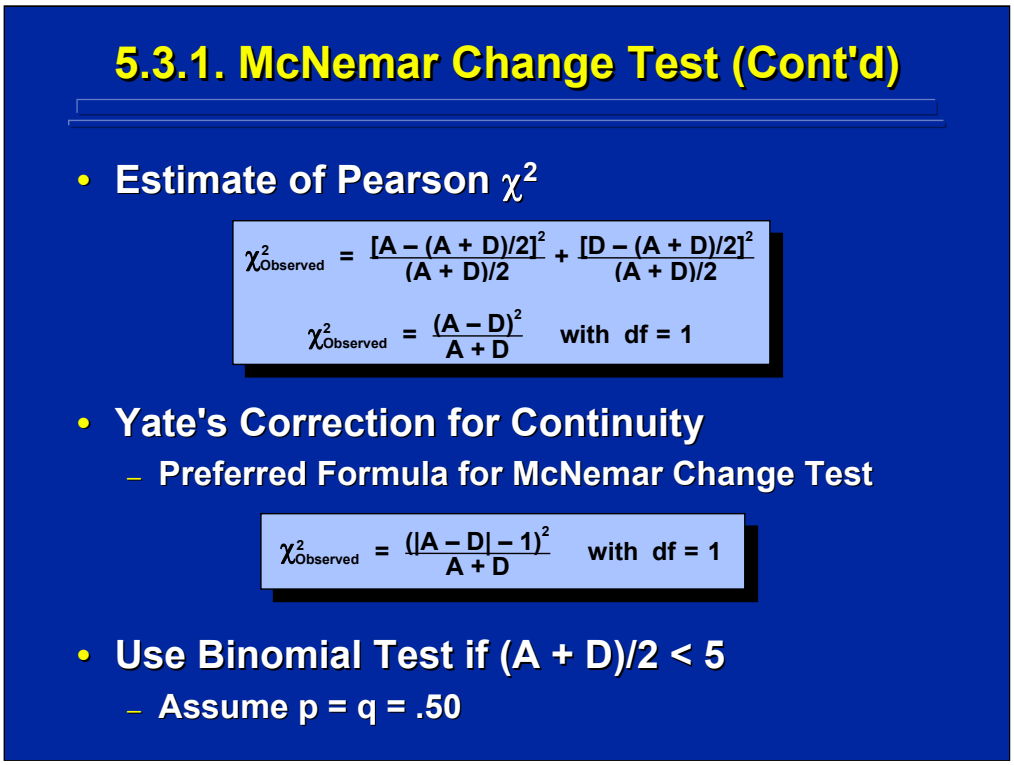
Although it is possible to determine the locus of the significant difference in larger contingency tables, to do so requires additional analyses based on meaningful partitions by the experimenter of the original contingency table. Consequently, evaluating the significant dependencies of the two classifications is not straightforward in chi-square tests of independence involving large contingency tables.




The two nonparametric tests for frequency data discussed are appropriate for within-subjects data. The McNemar Change Test is used for two categories, and the Cochran Q Test is used for more than two categories.




The McNemar Change Test is used primarily to evaluate changes in a subject's preference before and after use. Consequently, this is a within-subjects test based on two categories resulting in the 2x2 frequency table shown on this slide. Any change in preference is shown in cells A and D of the fourfold table. In cell A, the user goes from a positive preference before use to a negative preference after use; whereas, in cell D, the user goes from a negative preference beforehand to a positive preference after use. Consequently, A and D are the observed frequency of changes, and the average of A and D is the expected number of positive and negative changes.




The formula for the Pearson Chi-Square statistic based on observed and expected changes reduces to just A and D frequencies as shown on the top of this slide. Siegel and Castellan (1988, p. 76) recommend two alternatives to the McNemar test. First, the Yate's correction for continuity is the preferred formula for the McNemar test in order to provide a better approximation to the chi-square distribution. Second, the binomial test assuming equal probabilities instead of the McNemar test should be used if the expected frequency is less than 5 because the Pearson Chi-Square may not be distributed as chi-square in this circumstance.

## 5.3.1. McNemar Change Test (Cont'd)

- **Example Problem: 50 people stated their preference for Hearing Protectors A and B before and after using both protector on the job for one week at a time. Order of use was counterbalanced. Given the following data, did trial use of the hearing protectors change their preference (p < 0.05)?**

| | | After Using B | After Using A |
|---|---|---|---|
| Before Using | A | 26 | 13 |
| | B | 5 | 6 |

$$\chi^2_{Observed} = \frac{(|A - D| - 1)^2}{A + D} = \frac{(|26 - 6| - 1)^2}{26 + 6} = 11.28$$

$$\chi^2_{Tabled} = 1 \text{ df} = 3.84 \ (p < .05)$$

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows an example of using the McNemar Change Test to evaluate hearing protector preference. The experimenter wants to know if there is a significant difference in user preference before and after using each of two protectors. Since each user evaluates both hearing protectors, this requires a within-subjects test. The calculation of the Pearson Chi-Square using the Yate's correction is shown on this slide. Based on these results, there was a significant change in preference after use.

## 5.3.2. Cochran Q Test

- **Background**
  - Extension of McNemar Change Test to k ≥ 3 Matched Sets of Frequencies
  - Same Subjects or Matched Subjects
- **Use**
  - Compare Responses of "n" Subjects on "k" Conditions or Items
  - Dichotomous Response: "Success"=1 and "Failure"=0
- **Estimate of Pearson $\chi 2$ with (k-1) df**

$$Q = \frac{(k-1)\left[k\sum_{j=1}^{k} G_j^2 - \left(\sum_{j=1}^{k} G_j\right)^2\right]}{k\sum_{i=1}^{n} L_i - \sum_{i=1}^{n} L_i^2} \quad \text{and} \quad df = (k-1)$$

where,

$G_j$ = total number of "successes" in $j^{th}$ column

$L_i$ = total number of "successes" in $i^{th}$ row

If there are more than two related samples, k, the McNemar test can be extended to the Cochran Q test for 3 or more related samples. Usually related samples means the same subject is used in every condition; however, the Cochran Q test is also appropriate for closely matched subjects. Siegel and Castellan (1988, p. 173)) provide the formula shown on this slide for the observed Q statistic that is compared to the chi-square table value of k-1 degrees of freedom.

## 5.3.2. Cochran Q Test (Cont'd)

- **Example Problem: 15 experienced photo interpreters viewed a series of photographs under three enhancement procedures and rated each procedure as "acceptable - 1" or "unacceptable - 0". Are the three procedures rated equally acceptable ($p < 0.001$)?**

| Subject | Procedure 1 | Procedure 2 | Procedure 3 | $L_i$ | $L_i^2$ |
|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 1 | 1 |
| 2 | 1 | 1 | 1 | 3 | 9 |
| 3 | 0 | 1 | 1 | 2 | 4 |
| 4 | 0 | 1 | 1 | 2 | 4 |
| 5 | 0 | 1 | 1 | 2 | 4 |
| 6 | 0 | 1 | 1 | 2 | 4 |
| 7 | 1 | 1 | 1 | 3 | 9 |
| 8 | 0 | 1 | 1 | 2 | 4 |
| 9 | 0 | 0 | 1 | 1 | 1 |
| 10 | 0 | 1 | 1 | 2 | 4 |
| 11 | 1 | 1 | 1 | 3 | 9 |
| 12 | 0 | 1 | 1 | 2 | 4 |
| 13 | 1 | 1 | 1 | 3 | 9 |
| 14 | 0 | 1 | 1 | 2 | 4 |
| 15 | 0 | 1 | 1 | 2 | 4 |
| Total | $G_1 = 4$ | $G_2 = 14$ | $G_3 = 14$ | $\Sigma L_i = 32$ | $\Sigma L_i^2 = 74$ |

(Click in this red rectangle to see SAS calculations for this example.)

This is an example of using the Cochran Q test for evaluating the acceptability ratings of 15 photo interpreters of a series of photos using 3 different enhancement procedures. This is a within-subjects design since the same 15 photo interpreters evaluated each of the 3 photo enhancement procedures.

## 5.3.2. Cochran Q Test (Cont'd)

- **Calculations**

$$Q = \frac{(k-1)\left[k\sum_{j=1}^{k} G_j^2 - \left(\sum_{j=1}^{k} G_j\right)^2\right]}{k\sum_{i=1}^{n} L_i - \sum_{i=1}^{n} L_i^2}$$

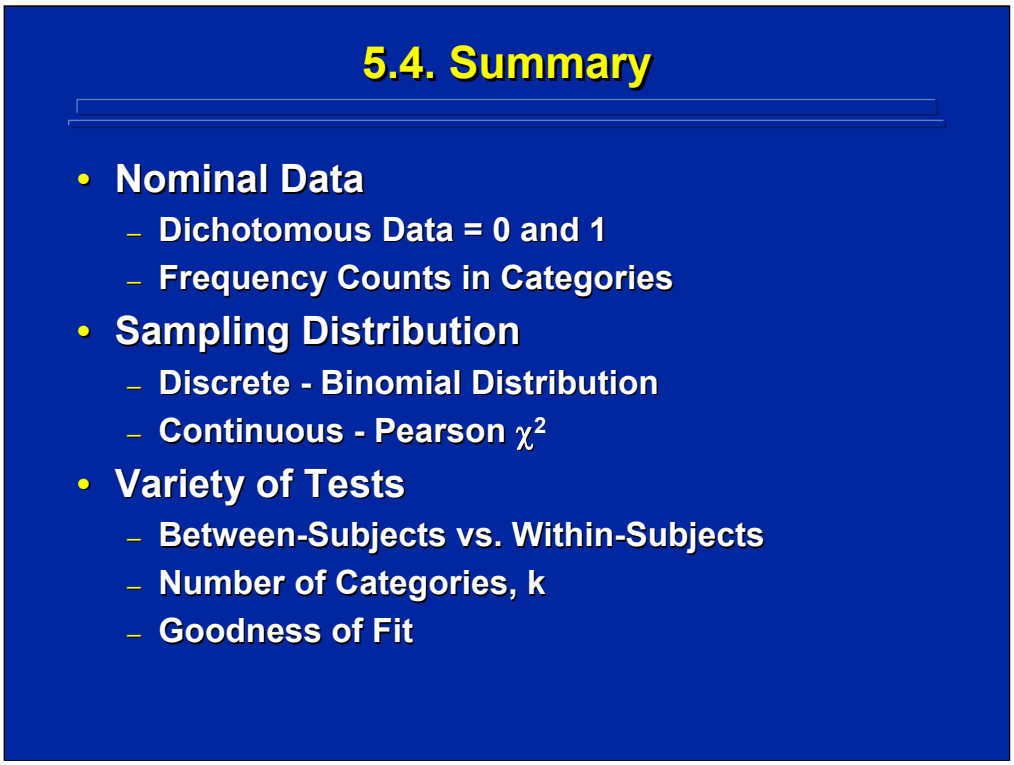
$$Q = \frac{(3-1)\left[3(4^2 + 14^2 + 14^2) - 32^2\right]}{3(32) - 74} = \boxed{18.18}$$

$$\chi^2_{Tabled} = (3-1) = 2 \text{ df} = \boxed{13.82} \quad (p < .001)$$

- **Paired Comparisons**
  - **Reduces to McNemar Change Test**

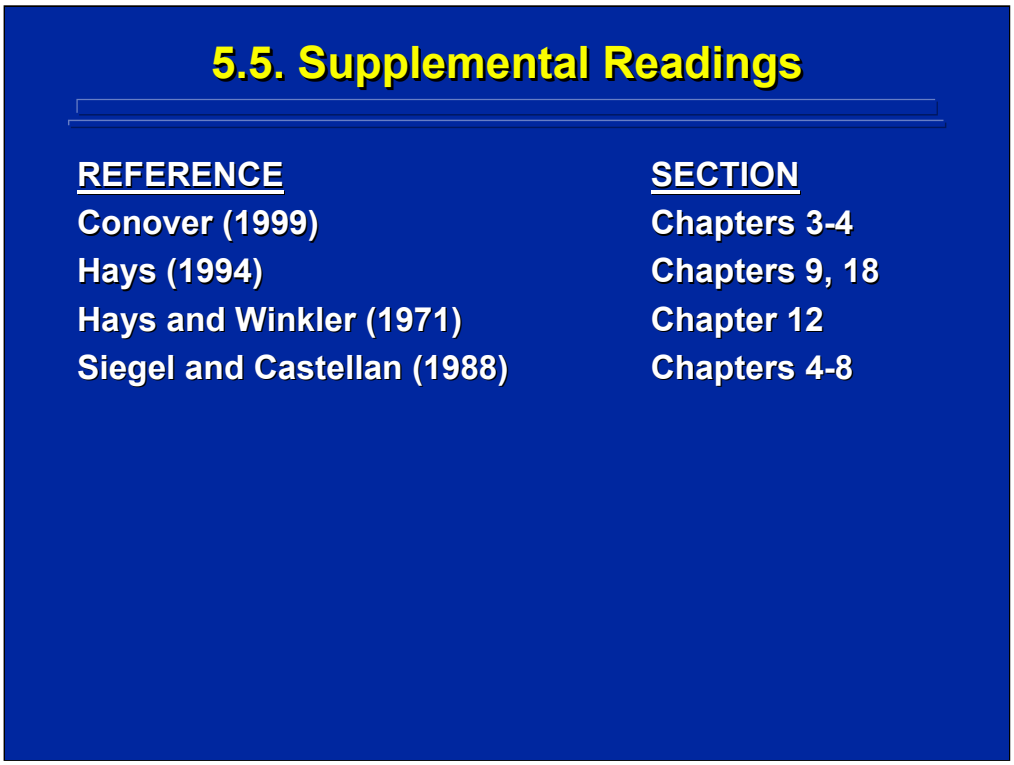
(Click in this red rectangle to see SAS calculations for this example.)

The calculations shown in this slide are based on the results shown on the previous data slide. After calculating the Q statistic, it is compared to a tabled chi-square of 2 degrees of freedom. Since the Q value is greater than the table value, one concludes that there is a significant difference of acceptability among the 3 photo enhancement procedures. In order to determine the locus of this difference, a series of subsequent McNemar Change Tests can be conducted on the paired comparisons of the 3 enhancement procedures.

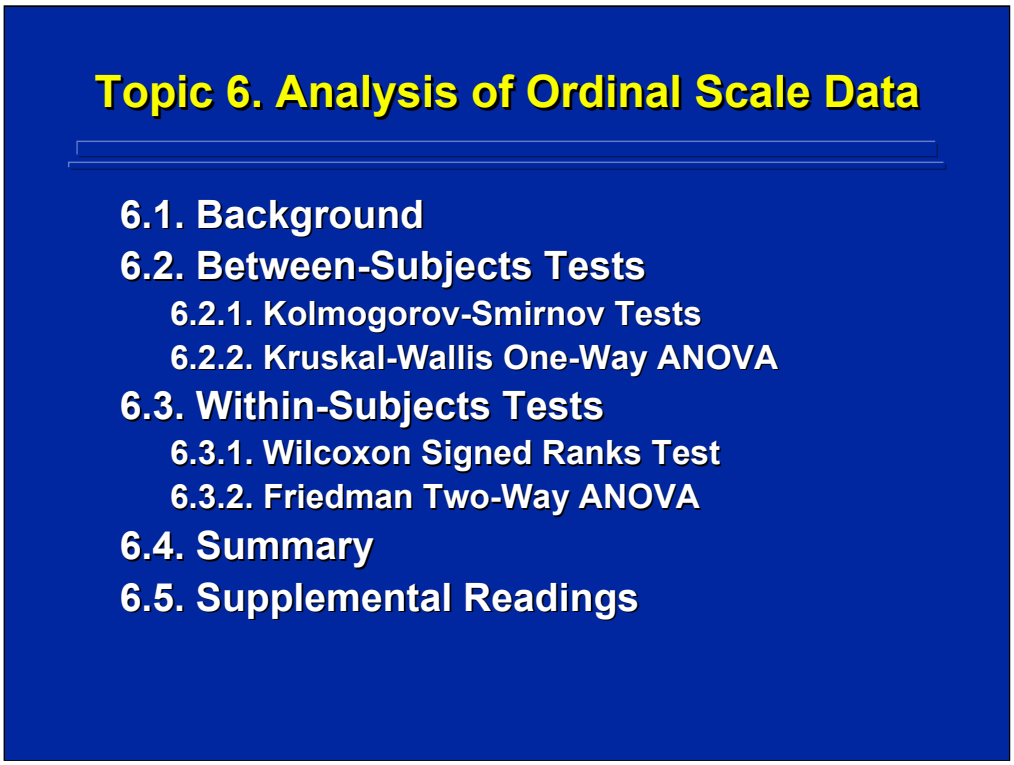



By way of summary, the four nonparametric procedures covered in this topic are representative of the most common procedures used in human factors and ergonomics research for analyzing nominal scale supplemental data. These data consist primarily of dichotomous data or frequency counts in categories. The Pearson chi-square statistic and the Yate's correction for continuity are used for these tests to approximate the continuous chi-square sampling distribution. To calculate the Pearson chi-square, one needs the observed value, O, from the data and an expected value, E, that is determined by the particular test used. A discrete binomial test can be used when E is less than 5.

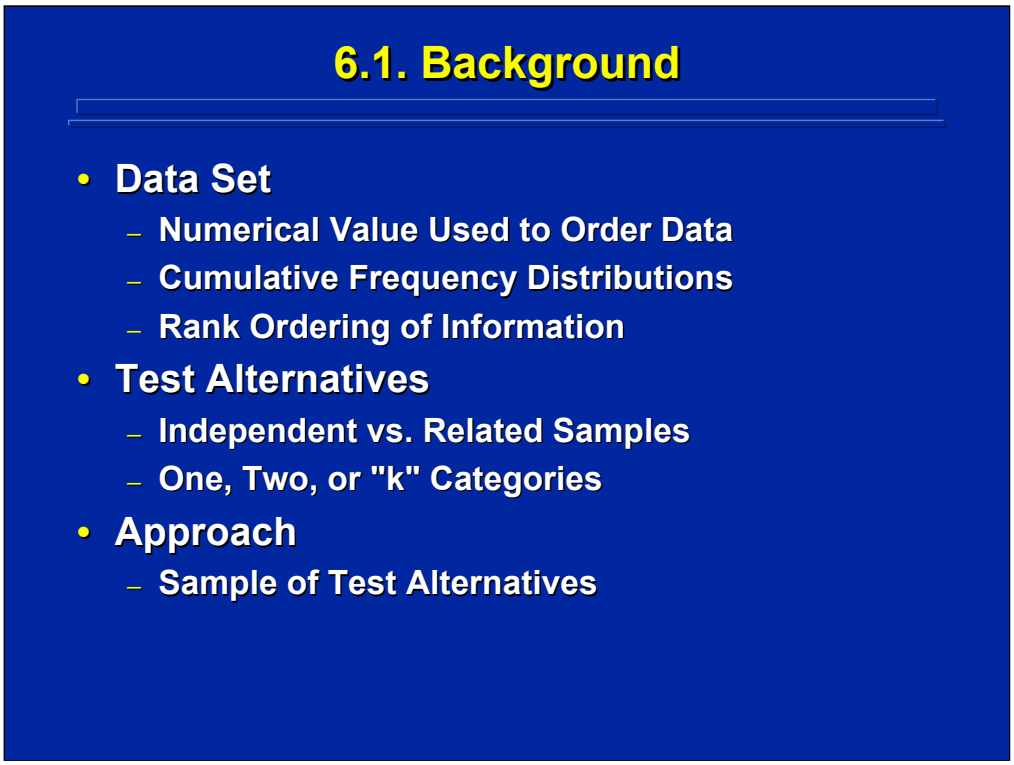
To choose the appropriate nominal data test, the experimenter first determines whether a between-subjects or within-subjects design is used. Next, the experimenter determines whether 2 or more categories, k, are to be included in the hypothesis test. A special case of two sample tests is the goodness of fit test in which the experimenter is comparing a single sample of nominal data to known population values.




Hays and Winkler (1971) and Hays (1994) provide an introductory overview of the Pearson Chi-Square, goodness of fit tests, and tests of independence. Siegel and Castellan (1988) is the classic nonparametric text used in behavioral research and human factors. All the formulae and tables as well as a more detailed discussion of the four nominal data analyses presented in this topic can be found in Siegel and Castellan (1988). Conover (1999) is another general reference on nonparametric analyses that provides further elaboration of the techniques covered in this reference topic.

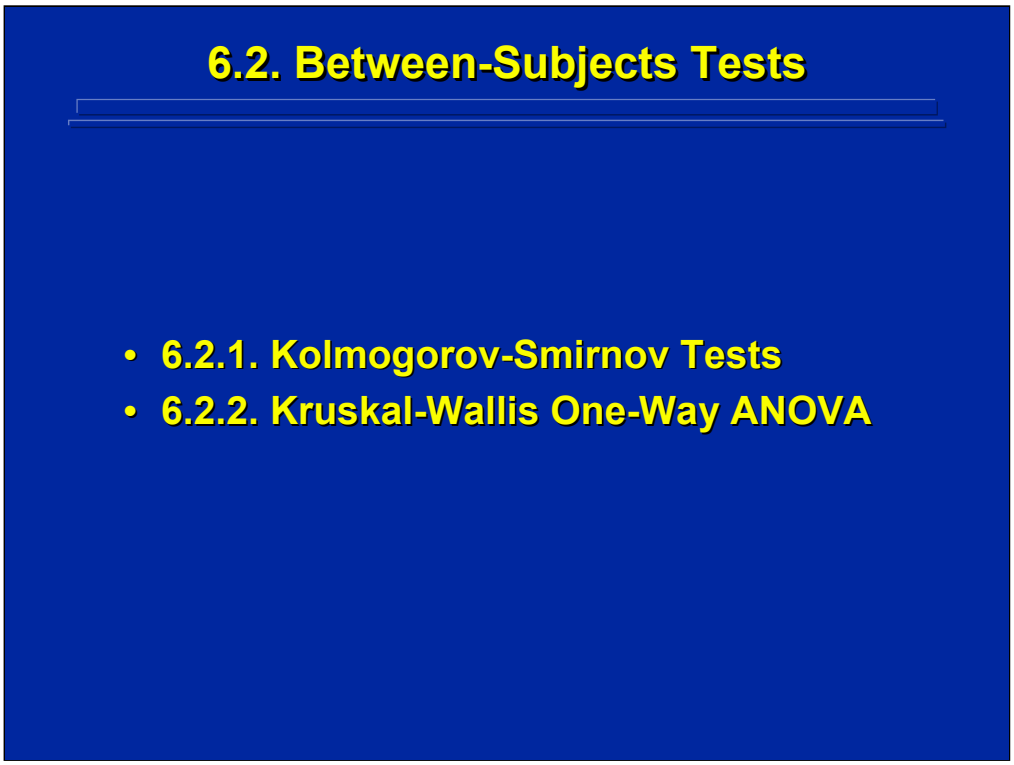



This topic deals with an overview of four nonparametric analysis alternatives that can be used with ordinal scale supplemental data. The four procedures described in this reference material are often used in human factors and ergonomics research. Once again, Siegel and Castellan (1988) provide a detailed discussion of each of these techniques, and their formulae and notation are used throughout this topic for easy reference. Similar to the approach followed in the discussion on nominal data analysis, the presentation of this topic is organized around between-subjects and within-subjects techniques to facilitate choice of nonparametric analysis procedures for ordinal data.

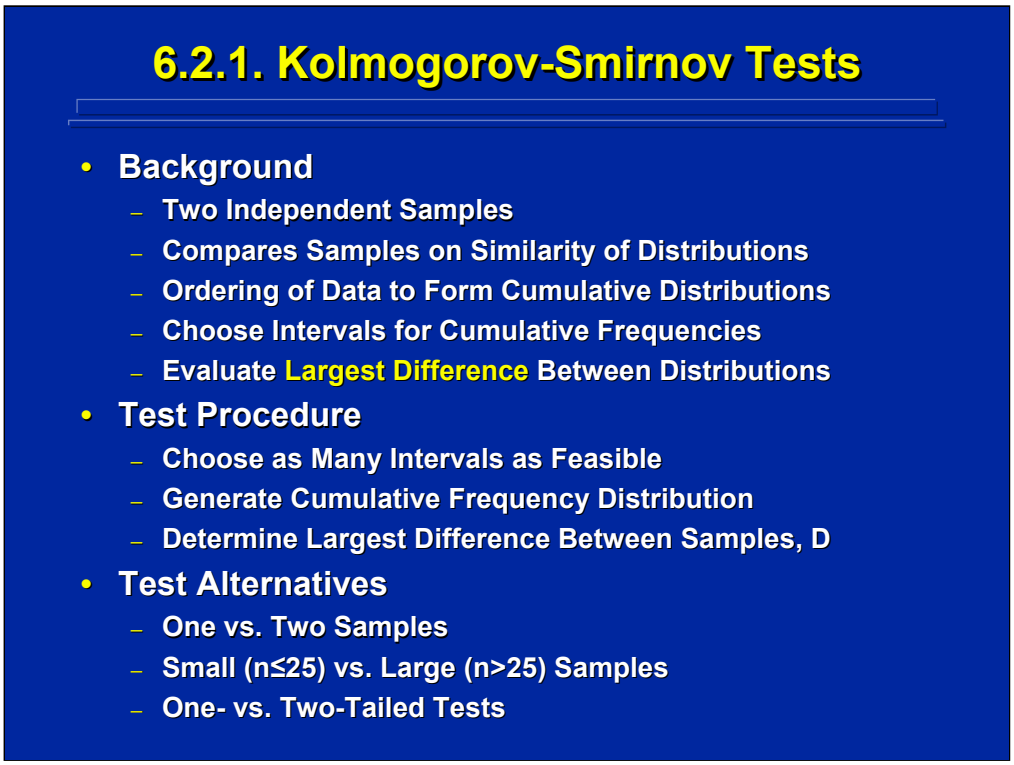



Ordinal data sets are numerical data in the form of cumulative frequencies or rank orders. Consequently, these data have order characteristics as well as frequency counts, or nominal characteristics. Ordinal data usually occur as cumulative frequency distributions or rank orders.

The choice of the appropriate ordinal nonparametric test depends upon whether the researcher has between-subjects or within-subjects samples and whether the researcher is testing one, two, or k categories. The four ordinal data analysis procedures discussed in this section cover a sample of the most common nonparametric procedures for between-subjects and within-subjects alternatives for a different number of categories.




Two of the most popular between-subjects tests of ordinal data in human factors are the Kolmogorov-Smirnov test and the Kruskal-Wallis One-Way ANOVA test. Choice between these two tests depends upon the number of categories being evaluated.

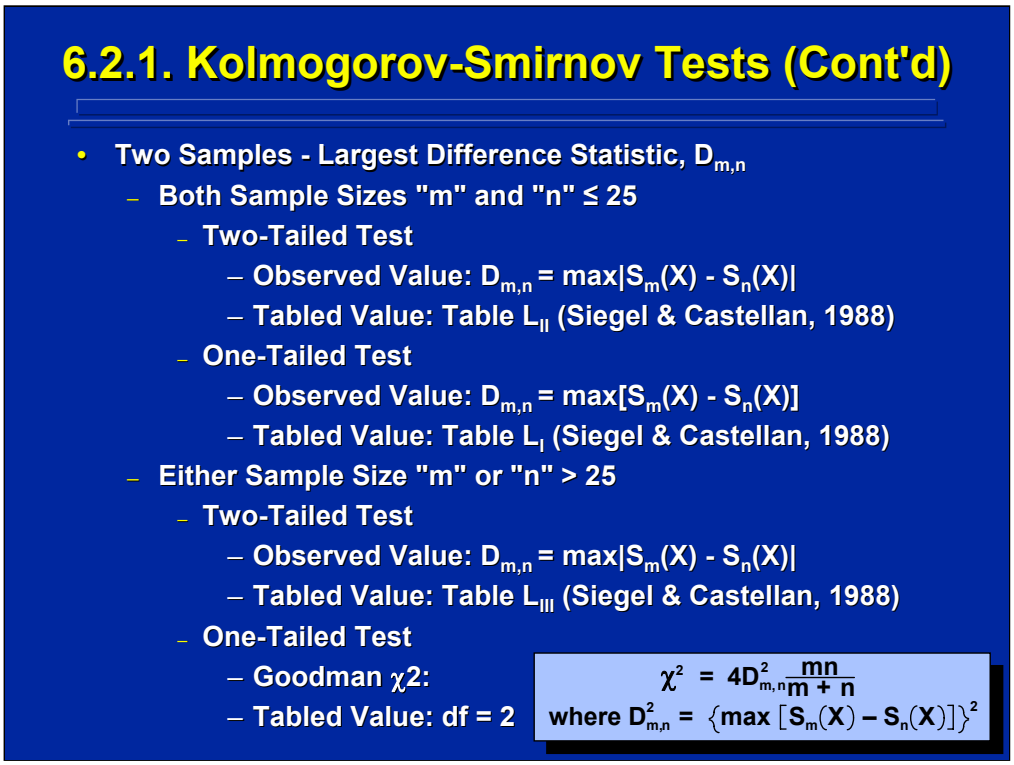



The Kolmogorov-Smirnov test was designed for two independent samples. It compares the similarities among the cumulative frequency distributions of samples. The test is based on the largest difference between the two cumulative distributions. The cumulative frequency distributions are based on meaningful intervals chosen by the experimenter. Various alternatives of the Kolmogorov-Smirnov test include one vs. two sample tests, small vs. large samples, and one tailed vs. two tailed tests.

## 6.2.1. Kolmogorov-Smirnov Tests (Cont'd)

- One Sample - Goodness of Fit Test
  - Observed Value
    - Maximum Deviation, $D_{Max}$
    - $D_{Max} = \max|F_0(X_i) - S_n(X_i)|$
      - where, i = 1, 2, ..., n
      - $F_0(X_i)$ = theoretical cumulative frequency distribution
      - $S_n(X_i)$ = observed cumulative frequency distribution of sample size, n
  - Tabled Value
    - Table F (Siegel & Castellan, 1988)
    - Based on Sample Size, n

A one sample test is goodness of fit test compares the sample cumulative frequency distribution to a known distribution. The observed value is the maximum absolute difference between the two cumulative frequency distributions as shown on this slide as defined by Siegel and Castellan (1988, p. 52). The tabled value is based on sample size and is presented in Table F in Siegel and Castellan (1988, p. 330).




The two-sample Kolmogorov-Smirnov test compares the observed largest difference, $D_{m,n}$, between cumulative frequency distributions of two independent samples where each sample can have different sample sizes, m and n, respectively. The observed value formulae for one-tailed versus two-tailed tests and small samples versus large samples are presented by Siegel and Castellan (1988, pp. 145-148). Depending upon sample size and choice of a one-tailed versus two-tailed test, the tabled values can be found in Siegel and Castellan (1988, pp. 348-352) Tables $L_I$, $L_{II}$, or $L_{III}$ as referenced on the slide.

If either of the two sample sizes is greater than 25, and the researcher is conducting a one-tailed test, then the Goodman chi-square approximation can be used to calculate the observed value according to the formula shown on this slide. This observed value is then compared to a tabled value from the chi-square sampling distribution based on 2 degrees of freedom.

## 6.2.1. Kolmogorov-Smirnov Tests (Cont'd)

- **Example Problem: 25 professional photographers and 30 nonprofessional photographers rated the "acceptability" of 25 photographs taken by an experimental camera on a 7-point Likert Scale. Median acceptability ratings of 25 photographs were determined for each individual. Did the nonprofessionals give significantly higher median ratings of acceptability ($p < 0.01$)?**

| Median Rating of Acceptability | Professional Photographers | Non-Professional Photographers |
|---|---|---|
| 1 | 9 | 1 |
| 2 | 6 | 3 |
| 3 | 1 | 2 |
| 4 | 2 | 5 |
| 5 | 4 | 8 |
| 6 | 2 | 7 |
| 7 | 1 | 4 |
| Total | 25 | 30 |

(Click in this red rectangle to see SAS calculations for this example.)

The example problem shown on this page is a one-tailed significance test based on two independent samples of cumulative frequency distributions of the 7 intervals of median acceptability ratings. Two preliminary analyses are required on each raw data set obtained from the 25 professional and 30 nonprofessional photographers. First, the median of the 7-point acceptability rating must be calculated across the 25 photographs for each subject in the two groups. Second, the frequency of each of the seven median rating values (i.e. 1 to 7) across subjects in each group determines the cumulative frequency distributions shown on this slide. Slater and Williges (2006) show the two raw data sets and median ratings for the data used in this example.

Only the data presented on this slide is needed to conduct the Kolmogorov-Smirnov Test. The two sample sizes of professional and non-professional photographers are different, and one of them is greater than 25. Consequently, the Goodman chi-square can be used to calculate the observed value in the subsequent hypothesis test using the chi-square sampling distribution.

## 6.2.1. Kolmogorov-Smirnov Tests (Cont'd)

- **Example Problem (Cont'd)**
  - **Cumulative Frequency Distributions**

| | Media Rating of Acceptability | | | | | | |
|---|---|---|---|---|---|---|---|
| Sample | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| $S_{25}(X)$ | 9/25 | 15/25 | 16/25 | 18/25 | 22/25 | 24/25 | 25/25 |
| $S_{30}(X)$ | 1/30 | 4/30 | 6/30 | 11/30 | 19/30 | 26/30 | 30/30 |
| $[S_{25}(X) - S_{30}(X)]$ | .327 | .467 | .440 | .353 | .247 | .093 | .000 |

  - **Observed Value**

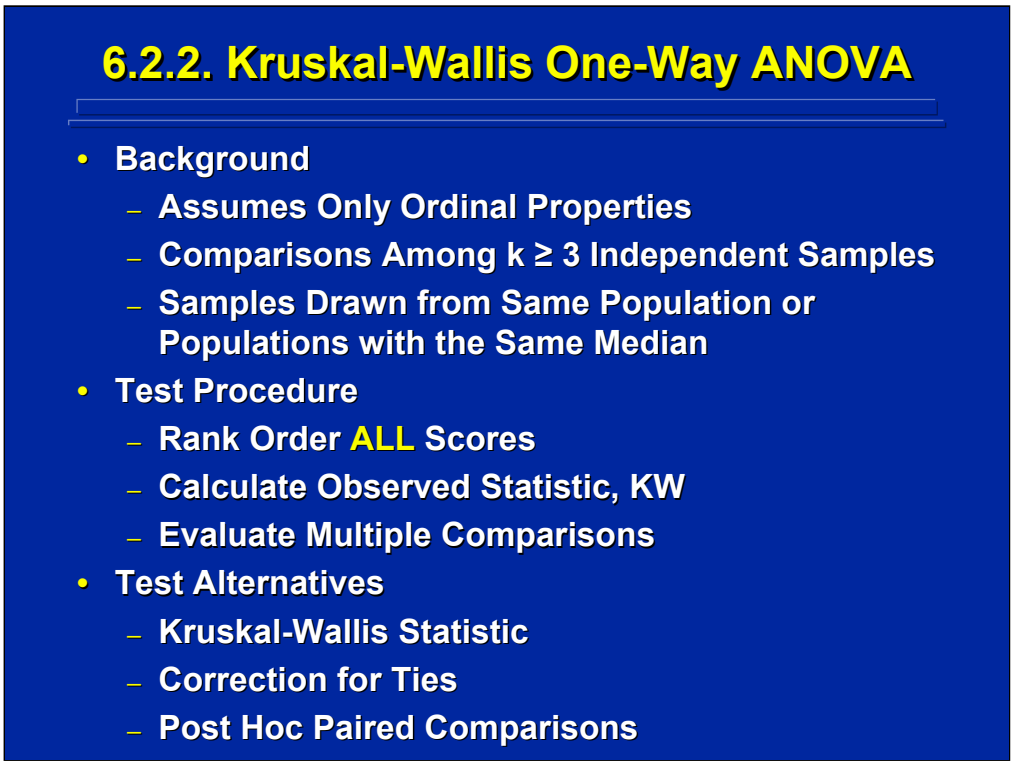
$$D^2_{25,30} = \{\max [S_{25}(X) - S_{30}(X)]\}^2 = (.467)^2$$

$$\chi^2 = 4D^2_{m,n}\frac{mn}{m+n} = \frac{4(.467)^2(25)(30)}{25+30} = 11.89$$

  - **Tabled Value: $\chi 2 = 9.21$ (2 df, $p < 0.01$)**
  - **Conclusion: Significantly Higher Acceptability Ratings by Nonprofessional Photographers**

(Click in this red rectangle to see SAS calculations for this example.)

Note that the top portion of this slide shows that the largest difference between the cumulative frequency distributions of acceptability ratings of the 25 professional and 30 nonprofessional photographers occurs at acceptability rating level 2 (i.e., 0.467). This interval is used to calculate the Goodman chi-square as shown on the middle portion of this slide. Since the observed value is greater than the tabled value, one concludes that the nonprofessional photographers had higher acceptability ratings than the professional photographers.




The Kruskal-Wallis test extends the Kolmogorov-Smirnov test to more than two independent samples (i.e., k>2). Since this analysis deals with 3 or more categories or levels of one factor, it is referred to as a one-way test of the factor of interest. The data used in the analysis have only ordinal properties.

Note that to conduct the Kruskal-Wallis statistic (KW) a rank order across all the scores in the entire data set is made before calculating the KW observed statistic. The KW calculation can be corrected for tied ranks. If a significant difference occurs, post hoc paired comparisons must be conducted to isolate the significant effect among the k samples since k is always greater than 2.

## 6.2.2. Kruskal-Wallis One-Way ANOVA (Cont'd)

- **Observed Value**
  - **Kruskal-Wallis Statistic, KW**

$$KW = \left[\frac{12}{N(N+1)}\sum_{j=1}^{k} n_j \bar{R}_j^2\right] - 3(N+1)$$

where, k = number of samples or groups
$n_j$ = number of cases in $j^{th}$ group
N = total number of observations
$\bar{R}_j$ = average of ranks in $j^{th}$ group

  - **Correction for Ties, $KW_T$**

$$KW_T = \frac{KW}{1 - \frac{\sum_{i=1}^{g}(t_i^3 - t_i)}{(N^3 - N)}}$$

where, g = number of groupings of different tied ranks
$t_i$ = number of tied ranks in the $i^{th}$ grouping
N = total number of observations

Both the general formula and the correction for ties for calculating the KW observed statistic as presented by Siegel and Castellan (1988, p. 207-210) are shown on this slide. Usually there is very little difference between the two calculations unless there are many tied ranks.

## 6.2.2. Kruskal-Wallis One-Way ANOVA (Cont'd)

- **Tabled Values**
  - Table O (Siegel & Castellan, 1988)
    - $k = 3$ when $n_1$ $n_2$ and $n_3 \leq 5$
  - $\chi 2$ Table with df = (k-1)
    - $k > 3$ or $n_j > 5$
- **Post Hoc Paired Comparisons**
  - Critical Difference between any Two Groups, U and V

$$|\bar{R}_U - \bar{R}_V| \geq Z_{\alpha/k(k-1)}\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_U} + \frac{1}{n_V}\right)}$$

  - Z Value - Tables A and $A_{II}$ (Siegel & Castellan, 1988)

The tabled value for the Kruskal-Wallis test is shown on the top portion of this slide. Table O from Siegel and Castellan (1988, p.356) can be used with 3 categories and small samples. If the number of categories is greater than 3 or the sample size is greater than 5, one can use the chi-squared table with k-1 degrees of freedom.

Since the Kruskal-Wallis test is used for 3 or more categories, a significant hypothesis test only tells the experimenter that at least one of the paired comparisons between categories is significant. One can use the unit normal sampling distribution to conduct subsequent paired comparisons to isolate the significant effect(s). The critical absolute difference formula for these paired comparisons according to Siegel and Castellan (1988, p. 213) is shown on the bottom of this slide where Tables A and $A_{II}$ in Siegel and Castellan (1988, pp. 319-320) can be used to determine the Z tabled value listed in the formula.

## 6.2.2. Kruskal-Wallis One-Way ANOVA (Cont'd)

- **Example Problem: A between-subjects design (n =6) was used to compare original learning by lecture, text, and multimedia instruction. Every trainee rated their overall satisfaction with the training on a 9-point scale. Did satisfaction differ across the three methods of training (p < .05)?**

| Satisfaction Ratings | | | Ranked Data | | |
|---|---|---|---|---|---|
| Lecture | Multimedia | Text | Lecture | Multimedia | Text |
| 7 | 7 | 3 | 13 | 13 | 3.5 |
| 4 | 9 | 6 | 5.5 | 17 | 10 |
| 5 | 9 | 5 | 7.5 | 17 | 7.5 |
| 6 | 6 | 4 | 10 | 10 | 5.5 |
| 7 | 8 | 2 | 13 | 15 | 2 |
| 3 | 9 | 1 | 3.5 | 17 | 1 |
| | | | $\bar{R}_j$ 8.75 | 14.83 | 4.92 |

(Click in this red rectangle to see SAS calculations for this example.)

Data from a hypothetical example problem comparing supplemental satisfaction ratings with three training techniques is shown on this slide. The Kruskal-Wallis One-Way ANOVA is appropriate for analyzing these satisfaction ratings, because a different group of subjects received each training method and 3 methods were compared.

Note that on the right hand portion of data table the overall rank order of all 18 satisfaction ratings across all three 3 training techniques is shown. Tied ranks are also shown. This resulting overall rank ordering is the raw data set used in the Kruskal-Wallis test.

## 6.2.2. Kruskal-Wallis One-Way ANOVA (Cont'd)

$$KW = \left[\frac{12}{N(N+1)}\sum_{j=1}^{k} n_j \bar{R}_j^2\right] - 3(N+1)$$

$$= \frac{12}{18(18+1)}[6(8.75)^2 + 6(14.83)^2 + 6(4.92)^2] - 3(18+1)$$

$$= \boxed{10.52}$$

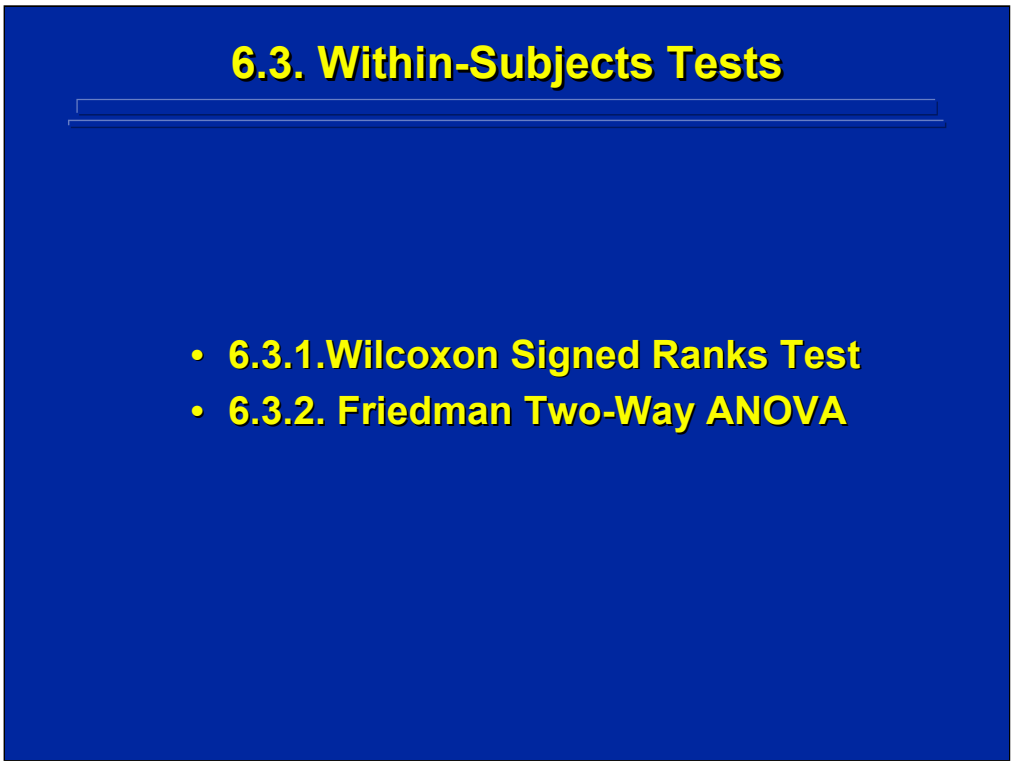
$$KW_T = \frac{KW}{1 - \frac{\sum_{i=1}^{g}(t_i^3 - t_i)}{(N^3 - N)}}$$

$$= \frac{10.52}{1 - \frac{3(2^3 - 2) + 3(3^3 - 3)}{(18^3 - 18)}} = \boxed{10.69}$$

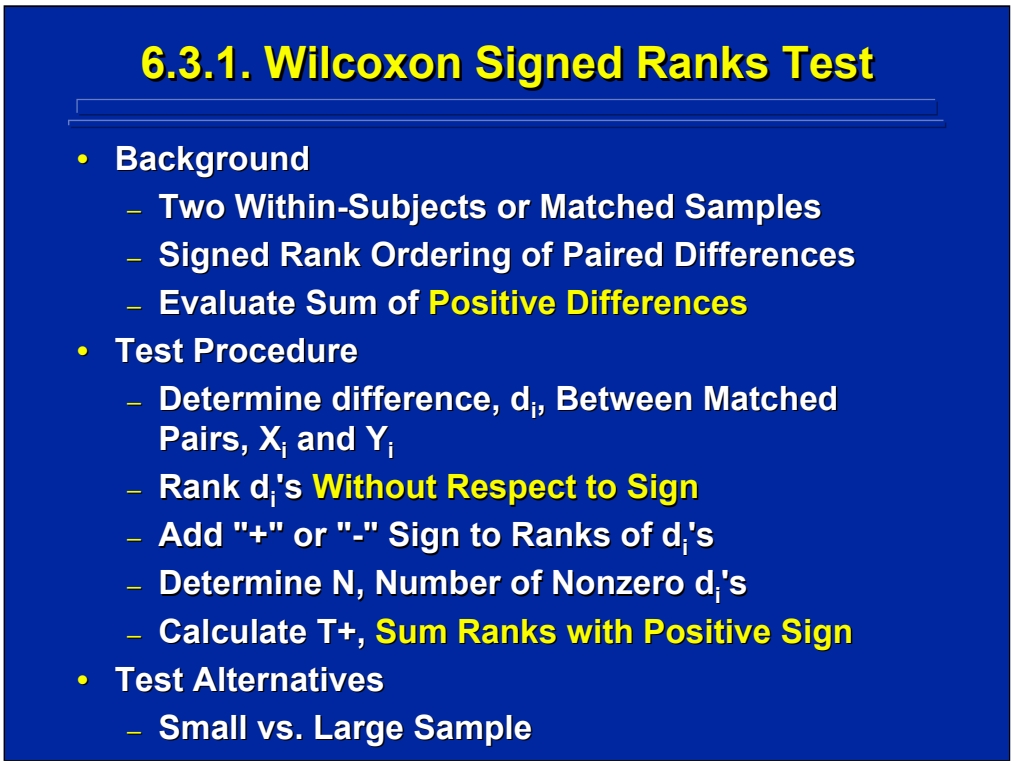
- Tabled Value
  - $\chi 2 = 5.99$ with 2 df ($p < 0.05$)
- Conclusion: Significant Difference

(Click in this red rectangle to see SAS calculations for this example.)

Two calculations for the observed KW statistic are shown on this slide. First the KW statistic is calculated assuming no tied ranks and, second, the correction for tied ranks is calculated since there are many ties in the dataset. Note that both calculations only differ slightly (10.52 and 10.69) and show a significant difference in satisfaction ratings among training techniques. Subsequent Z tests are needed to isolate these differences among the 3 training techniques.

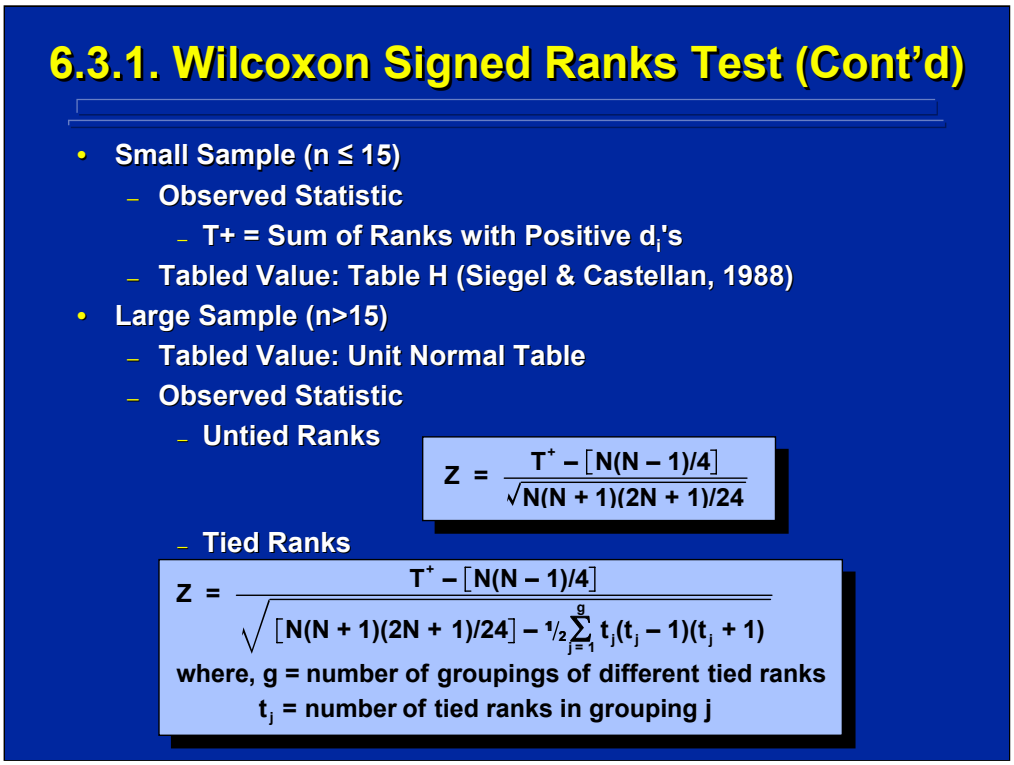



When the same subjects respond to every treatment category, within-subjects tests are needed. Two within-subjects tests are presented that are appropriate for ordinal data. The Wilcoxon Signed Ranks test which is used for two categories and the Friedman Two-Way ANOVA which is used for more than two categories.




The Wilcoxon test uses two within-subjects or matched subject samples. The test uses information about both the magnitude (i.e., rank order) as well as the direction of difference. The positive and negative differences between the two samples are determined and the test statistic is based on the sum of the rank order of only the positive differences. Hence, the name Signed Rank Test.

Procedurally, one first finds the differences between each pair of samples while maintaining the positive and negative relationships (i.e. the signed differences). Then one rank orders all differences without respect to sign where 1 is assigned to the smallest difference and so forth. Next one determines N, the number of nonzero differences. Finally, one calculates the T+ statistic, which is the sum of the ranks with a positive sign. Test alternatives vary depending on whether one has small or large tests.




For small samples of 15 or less, one would use the T+ statistic (i.e., the sum of all positive ranks) as the observed statistic. The table value of T+ is provided in Table H in Siegel and Castellan (1988, pp. 332-334).

For samples larger than 15, the unit normal sampling distribution can be used to determine the tabled value. The Z observed formula for both untied and tied positive ranks according to Siegel and Castellan (1988, pp. 91, 94) is presented on this slide.

## 6.3.1. Wilcoxon Signed Ranks Test (Cont'd)

- **Example Problem: Two electronic communication methods, video conferencing and instant messaging, were evaluated by each of 11 soldiers in a battlefield information setting on four 9 point Likert-Type Scales in terms of ease of use, effectiveness, timeliness, and convenience. Are the two communication methods significantly different in terms of overall acceptability as measured by the sum of these four ratings ($p < 0.05$)?**
- **N = 11**
- **T+ = (2+6+8+11+7+5+9+10) = 58**
- **Table H (Siegel & Castellan, 1988) = $p < 0.0244$ (Two-Tailed)**

| Soldiers | Conferencing | Messaging | d | Rank of d |
|---|---|---|---|---|
| 1 | 29 | 26 | 3 | 2 |
| 2 | 17 | 11 | 6 | 6 |
| 3 | 8 | 12 | -4 | -3.5 |
| 4 | 21 | 8 | 13 | 8 |
| 5 | 33 | 5 | 28 | 11 |
| 6 | 30 | 19 | 11 | 7 |
| 7 | 25 | 20 | 5 | 5 |
| 8 | 24 | 10 | 14 | 9 |
| 9 | 15 | 19 | -4 | -3.5 |
| 10 | 10 | 12 | -2 | -1 |
| 11 | 34 | 18 | 16 | 10 |

(Click in this red rectangle to see SAS calculations for this example.)

This is an example using a Wilcoxon Signed Ranks Test. The example problem shown on this slide provides the sum of 4 acceptability ratings obtained from 11 soldiers after they use each of two types of electronic communication devices. The significant difference of acceptability between the two communication methods can be determined by the Wilcoxon Signed Ranks Test for small samples.

Signed acceptability differences between the two communication systems are shown in the "d" column of the slide. The T+ statistic, based on sum of all positive ranks shown in the right most column of the slide, provides the observed value that is compared to the tabled value found in Table H from Siegel and Castellan (1988, p. 333). One can conclude that the overall acceptability rating as measured by the sum of the four sub-ratings is significantly different for the two electronic communication methods at the 0.05 level of significance.

## 6.3.1 Wilcoxon Signed Ranks Test (Cont'd)

- **Sub-Ratings of Acceptability**

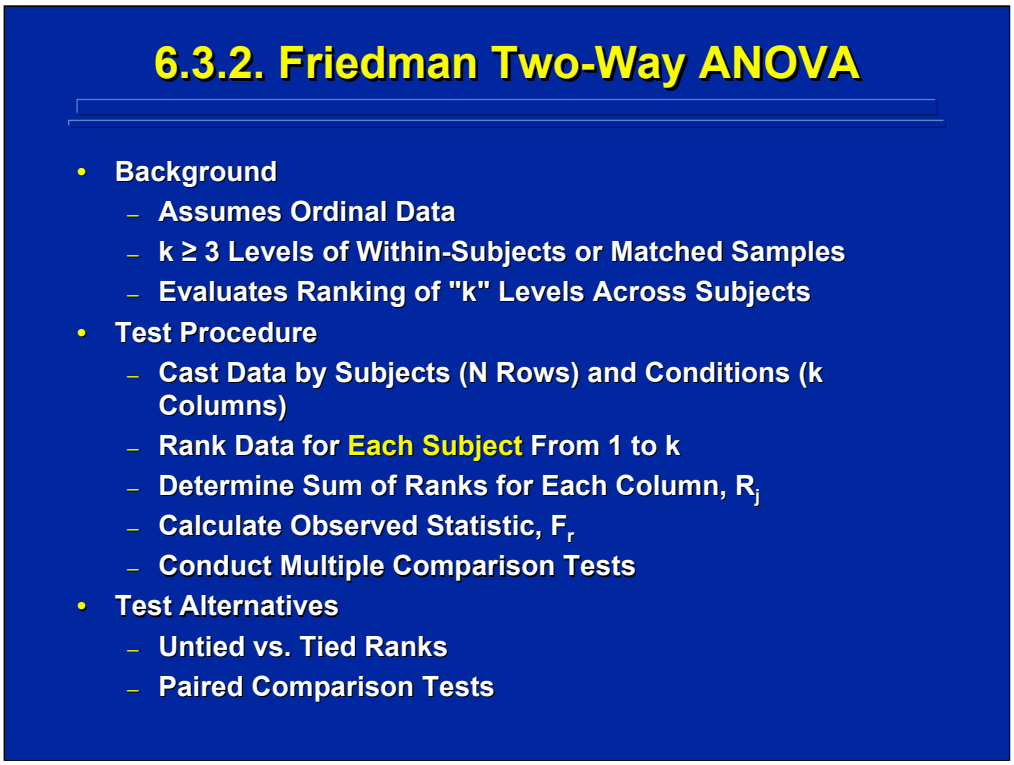
| Video Conferencing | | | | | |
|---|---|---|---|---|---|
| Soldier | Ease of Use | Effectiveness | Timeliness | Convenience | Sum |
| 1 | 8 | 7 | 8 | 6 | 29 |
| 2 | 4 | 3 | 6 | 4 | 17 |
| 3 | 2 | 1 | 3 | 2 | 8 |
| 4 | 5 | 2 | 6 | 8 | 21 |
| 5 | 9 | 8 | 7 | 9 | 33 |
| 6 | 7 | 6 | 8 | 9 | 30 |
| 7 | 6 | 4 | 9 | 6 | 25 |
| 8 | 5 | 6 | 7 | 6 | 24 |
| 9 | 4 | 1 | 4 | 6 | 15 |
| 10 | 3 | 2 | 4 | 1 | 10 |
| 11 | 9 | 8 | 9 | 8 | 34 |

| Instant Messaging | | | | | |
|---|---|---|---|---|---|
| Soldier | Ease of Use | Effectiveness | Timeliness | Convenience | Sum |
| 1 | 7 | 8 | 5 | 6 | 26 |
| 2 | 4 | 5 | 1 | 1 | 11 |
| 3 | 3 | 5 | 3 | 1 | 12 |
| 4 | 2 | 2 | 2 | 2 | 8 |
| 5 | 2 | 1 | 1 | 1 | 5 |
| 6 | 5 | 6 | 4 | 4 | 19 |
| 7 | 5 | 3 | 5 | 7 | 20 |
| 8 | 4 | 2 | 2 | 2 | 10 |
| 9 | 4 | 3 | 6 | 6 | 19 |
| 10 | 1 | 2 | 4 | 5 | 12 |
| 11 | 6 | 4 | 3 | 5 | 18 |

(Click in this red rectangle to see SAS calculations for this example.)

Note the Wilcoxon Signed Ranks Test used in this example was based on the "overall acceptability" as measured by the sum of the 4 sub-ratings shown on this slide. Consequently, conclusions can only be made in terms of overall acceptability of the video conferencing and instant messaging communication systems.

Additional analyses would be required if one were interested in drawing separate conclusions about ease of use, effectiveness, timeliness, and convenience. Four additional Wilcoxon Signed Ranks Tests could be conducted, one on each of the separate sub-rating scale results shown on this slide to isolate components of the significant overall acceptability rating of the 11 soldiers.




The Friedman Two-Way ANOVA is appropriate for ordinal data representing more than 2 categories collected from within-subjects or matched samples. The data set is organized by subjects in “N” rows, and by levels in “k” columns. Hence, this test is referred to as a two-way test.

Procedurally, one rank orders the data for each subject separately from 1 to k for each column. Then one determines the sum of the ranks for each column, $R_j$. Next one calculates the observed statistic, $F_r$, using either the formula for tied or untied ranks. Finally, one conducts multiple comparison tests to isolate significant effects.

## 6.3.2. Friedman Two-Way ANOVA (Cont'd)

- **Observed Statistic, $F_r$**
  - **Untied Ranks**

$$F_r = \left[\frac{12}{Nk(k+1)}\sum_{j=1}^{k} R_j^2\right] - 3N(k+1)$$

  - **Tied Ranks**

$$F_r = \frac{12\sum_{j=1}^{k} R_j^2 - 3N^2k(k+1)^2}{Nk(k+1) + \frac{\left(Nk - \sum_{i=1}^{N}\sum_{j=1}^{g_i} t_{i,j}^3\right)}{(k-1)}}$$

where, $g_i$ = number of sets of tied ranks in $i^{th}$ group
$t_{i,j}$ = size of $j^{th}$ set of tied ranks in $i^{th}$ group

This slide shows the formulae presented by Siegel and Castellan (1988, pp. 177, 179) for calculating the $F_r$ observed statistic for either tied or untied ranks. Usually there is little difference in the result of each formula unless there are many tied ranks. Note both formulae are based on the number of subject, N, the number of columns (conditions), k, and the sum of the ranks for each column, $R_j$.

## 6.3.2. Friedman Two-Way ANOVA (Cont'd)

- Tabled Values
  - Small Sample (k≤5 and N in Table)
    - Table M (Siegel & Castellan, 1988)
  - Large Sample (k>5 or N not in Table)
    - $\chi^2$ with (k-1) df
- Post Hoc Paired Comparisons
  - Critical Difference between any Two Groups, U and V
    - Total of Each Ranking $|R_U - R_V| \geq Z_{\alpha/k(k-1)}\sqrt{\frac{Nk(k+1)}{6}}$
    - Average of Each Ranking $|\bar{R}_U - \bar{R}_V| \geq Z_{\alpha/k(k-1)}\sqrt{\frac{k(k+1)}{6N}}$
  - Z Value – Tables A and $A_{II}$ (Siegel & Castellan, 1988)

Siegel and Castellan (1988, p. 353) provide tabled values for small sample sizes in Table M. The chi-square sampling distribution can be used for samples greater than five.

If the resulting Friedman test is significant, the experimenter knows that at least one of the paired comparisons between treatment levels is significant. Post hoc paired-comparisons are needed to isolate these differences. Subsequent post hoc comparisons can be conducted that are based on the unit normal sampling distribution. The critical absolute difference formulae for these paired comparisons for either totals or means according to Siegel and Castellan (1988, p. 180) is shown on the bottom of this slide where Tables A and $A_{II}$ in Siegel and Castellan (1988, pp. 319-320) can be used to determine the Z tabled value listed in the formulae.

## 6.3.2. Friedman Two-Way ANOVA (Cont'd)

- **Example Problem**: Five Subjects performed a benchmark task using a new CAD program. After completing the task, users rated their satisfaction using QUIS, and median ratings were calculated for each of the four parts of the scale, i.e., I. Screen, II. Terminology, III. Learning, and IV. Capability. Did median satisfaction differ across the parts ($p < 0.05$)?
- $F_r = [12/(5)(4)(4+1)][7^2+17^2+16^2+10^2] - (3)(5)(4+1) = 8.28$
- Table M (N=5, k=4, $p < 0.05$) = 7.8

**Median Ratings of CAD Usability Parts of QUIS**

| Subjects | I | II | III | IV |
|---|---|---|---|---|
| 1 | 2 | 6 | 7 | 3 |
| 2 | 4 | 8 | 9 | 3 |
| 3 | 1 | 9 | 6 | 2 |
| 4 | 0 | 5 | 8 | 1 |
| 5 | 5 | 7 | 4 | 6 |

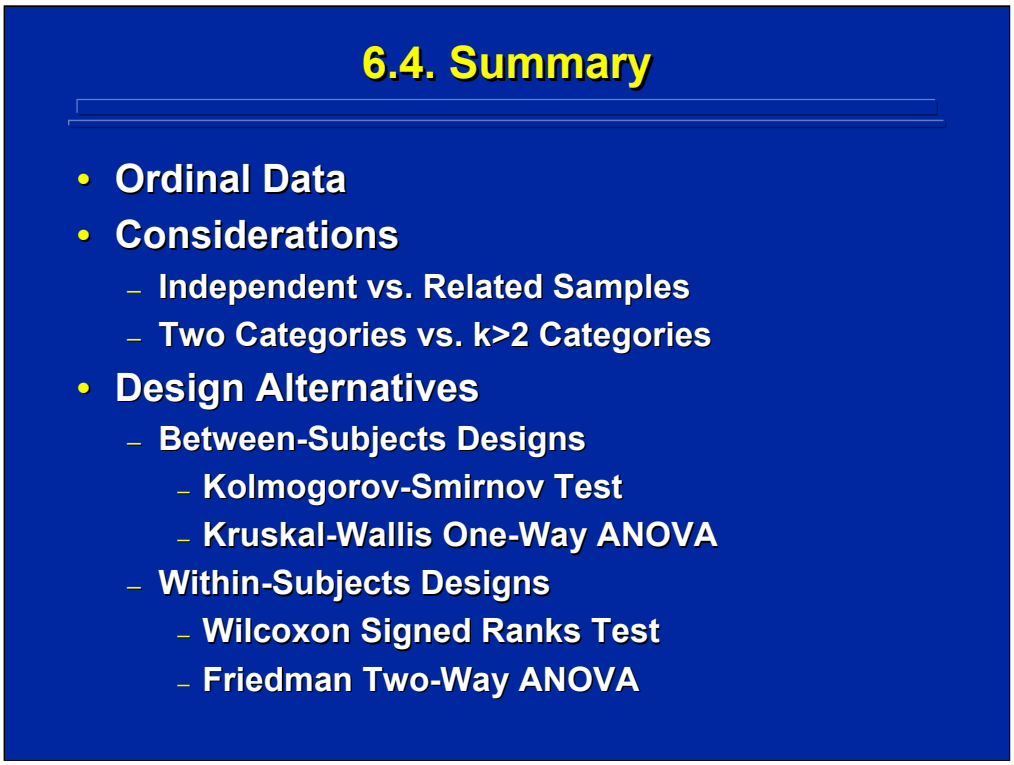
**Rankings for $F_r$ Calculations Parts of QUIS**

| Subjects | I | II | III | IV |
|---|---|---|---|---|
| 1 | 1 | 3 | 4 | 2 |
| 2 | 2 | 3 | 4 | 1 |
| 3 | 1 | 4 | 3 | 2 |
| 4 | 1 | 3 | 4 | 2 |
| 5 | 2 | 4 | 1 | 3 |
| Rj | 7 | 17 | 16 | 10 |

(Click in this red rectangle to see SAS calculations for this example.)

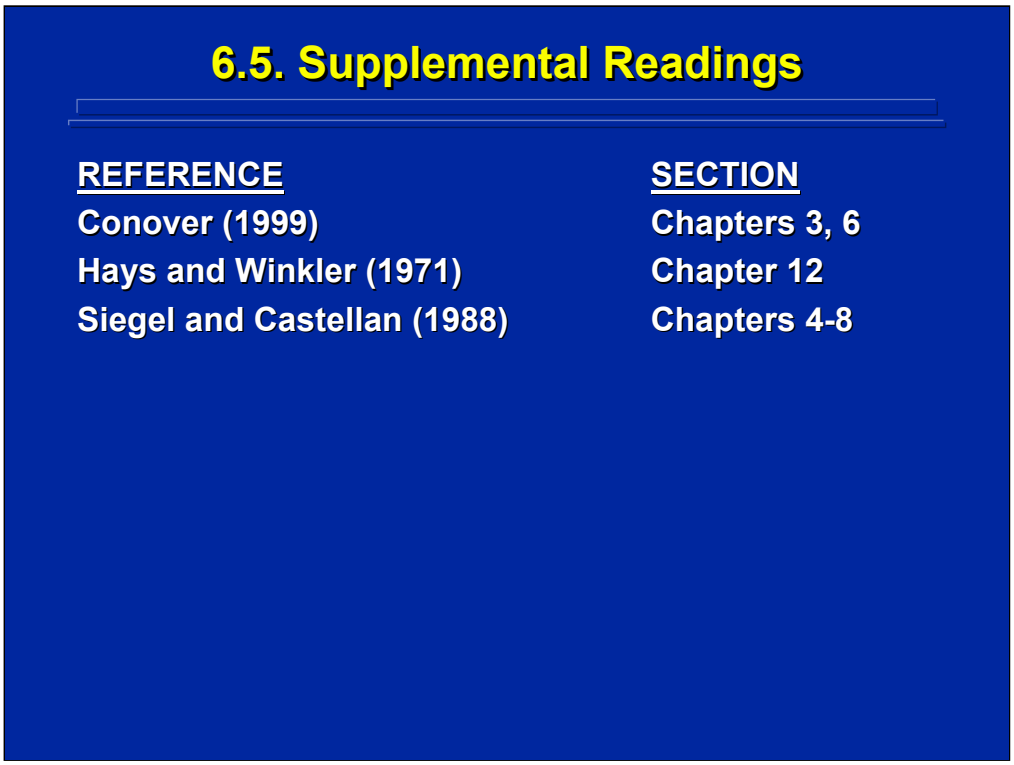
In this example, the satisfaction of 5 subjects was evaluated on each of the 4 parts of the QUIS scale after using a CAD system. The Friedman Test can be used to determine if the subjects' median ratings of satisfaction as shown in the left hand side of the data set on this slide differs significantly across the 4 parts of QUIS.

Note that a rank order of the median ratings is made for each subject separately across the 4 parts of QUIS as shown in the right hand side of the data set shown on the slide. These rank orderings are used to calculate the observed value, $F_r$, that is then compared to the tabled value. As shown on the slide, there is a significant difference in user satisfaction ratings across various parts of QUIS since the observed value, 8.28, is greater than the tabled value, 7.8. Subsequent post hoc paired-comparison tests are needed to isolate these differences.

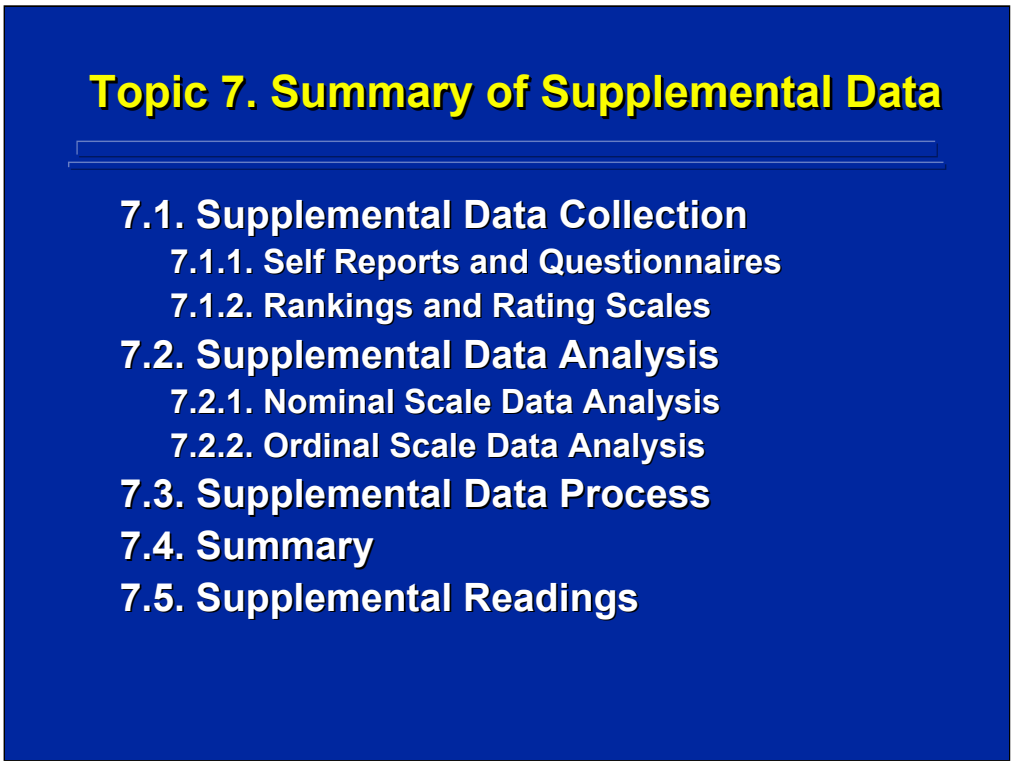



By way of summary, the four procedures described in this reference topic are used to test significant differences in supplemental data that have ordinal characteristics. Ordered data usually appear as cumulative frequencies or rank orders. Remember that researchers must first determine if they are comparing just 2 categories or more than 2 categories in choosing the appropriate significant test. Next researchers need to know if they have between-subjects or within-subjects data.

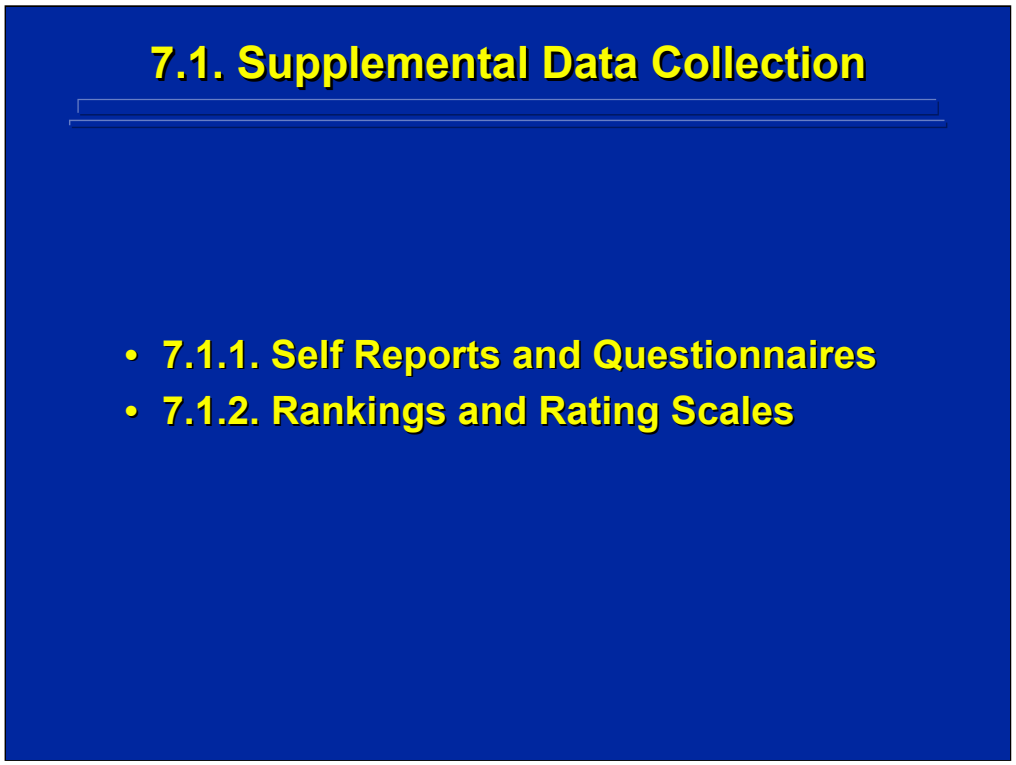
As shown on this slide, the resulting choice of testing procedure is rather straightforward. The design alternatives for between-subjects data are either the Kolmogorov-Smirnov test (2 categories), or the Kruskal-Wallis One-Way ANOVA test (more than 2 categories). The alternatives for within-subjects data are either the Wilcoxon Signed Ranks test (2 categories) or the Friedman Two-Way ANOVA (more than 2 categories).



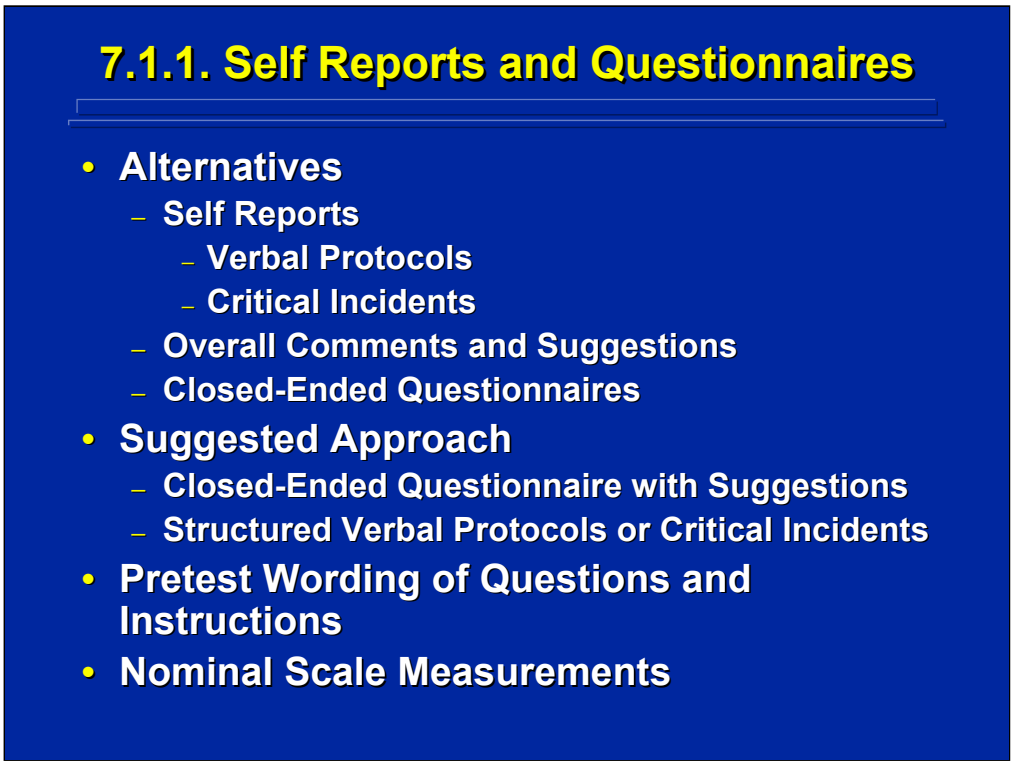

Hays and Winkler (1971) provide an introductory overview of rank order tests. Siegel and Castellan (1988) provides the most complete coverage of the ordinal data procedures covered in this topic. All the formulae and a more detailed discussion of the 4 ordinal data analysis procedures and tables presented in this topic can be found in Siegel and Castellan (1988). Conover (1999) is another general reference on nonparametric analyses that provides further elaboration of the techniques covered in this reference topic.




This topic summarizes the major points discussed in Section 2 dealing with supplemental data. First, supplemental data collection procedures dealing with self reports, questionnaires, and rating scales are reviewed. Second, supplemental data analysis techniques using common nominal and ordinal scale nonparametric procedures are reviewed. Third, a three-step process for dealing with supplemental data in experimental design is presented. Finally, an overall summary of this process and appropriate supplemental readings provided in Section 2 of this reference material are listed.

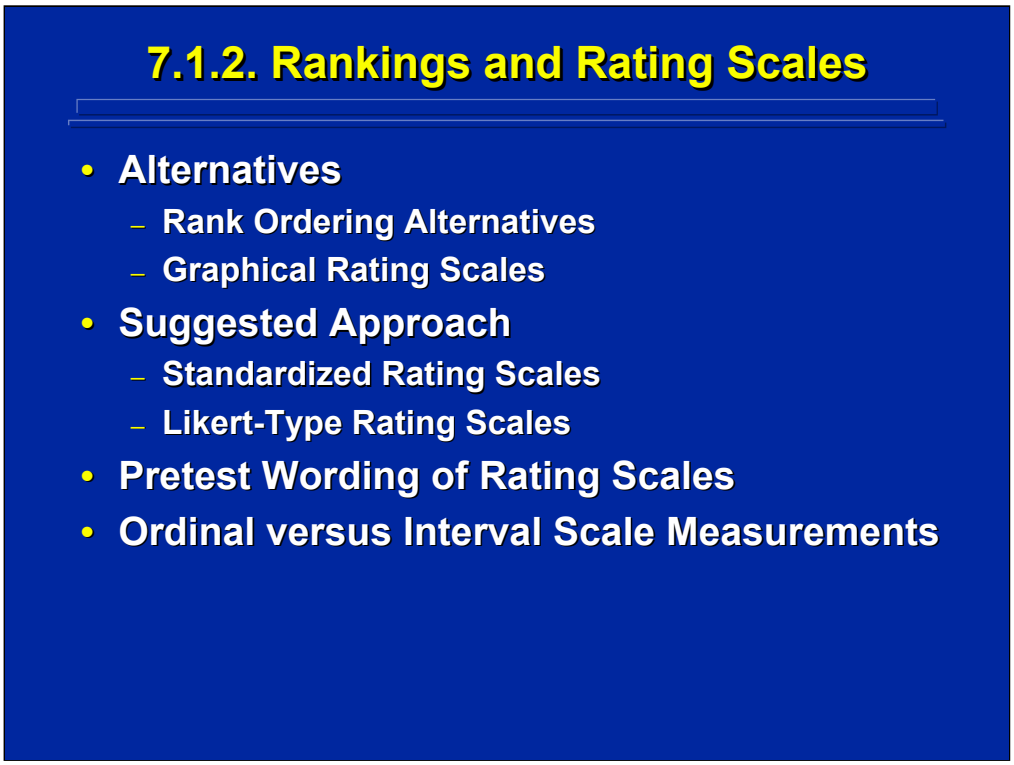



Supplemental data collection consists of subjective opinions and demographic data collected in addition to the main dependent variables measured in the experimental design in order to aid in the interpretation of the main results of the experiment. The experimenter should attempt to collect objective supplemental data that is quantitative to facilitate subsequent analysis and interpretation. Both self reports and questionnaires are forms of supplemental data obtained from subjects participating in the experiment. In addition, various forms of rankings and ratings can be used.



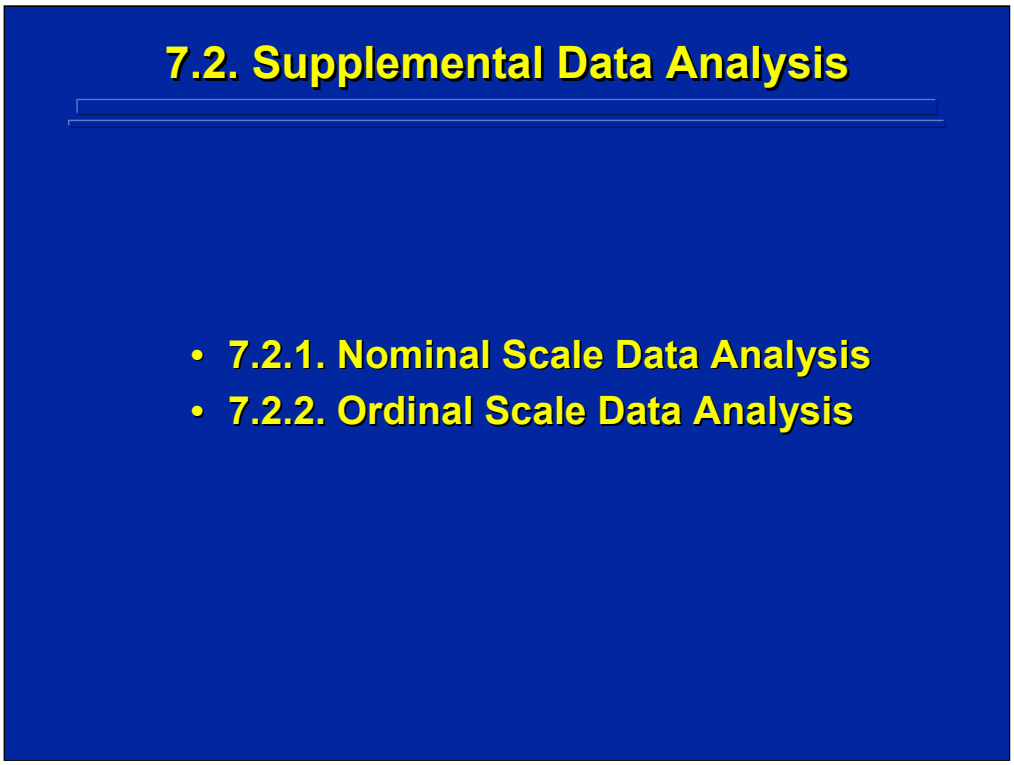

Details on self reports and questionnaires are presented in Topic 4 of this reference material. Several alternative methods such as verbal protocols, critical incidents, and closed end questionnaires are suitable for subsequent quantitative analysis of the supplemental data. Overall comments and suggestions provided by subjects in an experiment are usually tabulated. Researchers often use a closed-ended questionnaire followed by an open-ended question for overall comments from subjects as the major form of supplemental data collection from subjects in a human factors experiment. When detailed self-reports from subjects are needed throughout the experiment, human factors researcher often use structured verbal protocols and critical incidents methods to collect these data.

Pretesting the wording and instructions is critical before using any self report or questionnaire in an experiment in order to avoid confusion and unreliable results during data collection. Since self reports and questionnaires result in frequency counts and tabulations, nonparametric analyses of these nominal scale measurements are appropriate.

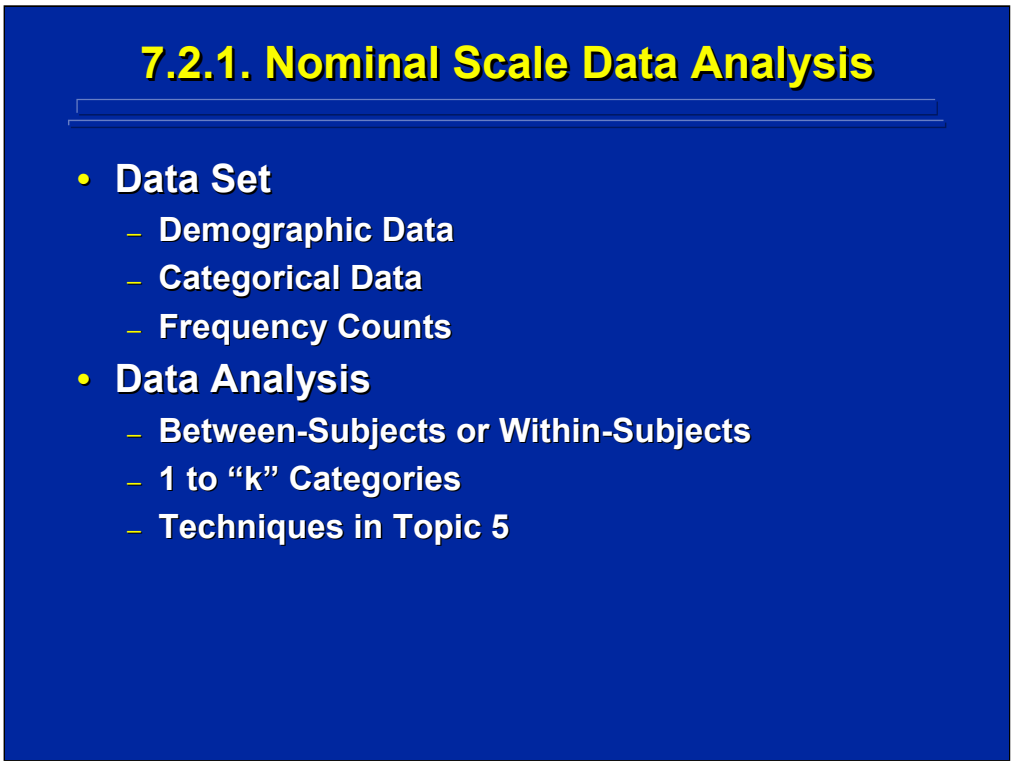



Often the experimenter asks subjects to rank order preferences or use a graphical rating scales as a systematic means of obtaining supplemental data from the subjects. Standardized rating scales and Likert-type rating scales as described in Topic 4 are most often used by human factors researchers. Pretesting the wording used in rating scales and rankings are critical to collecting reliable and valid supplemental data.

Ratings and rank orders provide ordinal scale measurements that are amenable to subsequent nonparametric statistical analysis. Although some human factors researchers assume that Likert-type rating scales have the properties of interval scale measurement and use a parametric statistical analysis on those data, this is usually not appropriate.

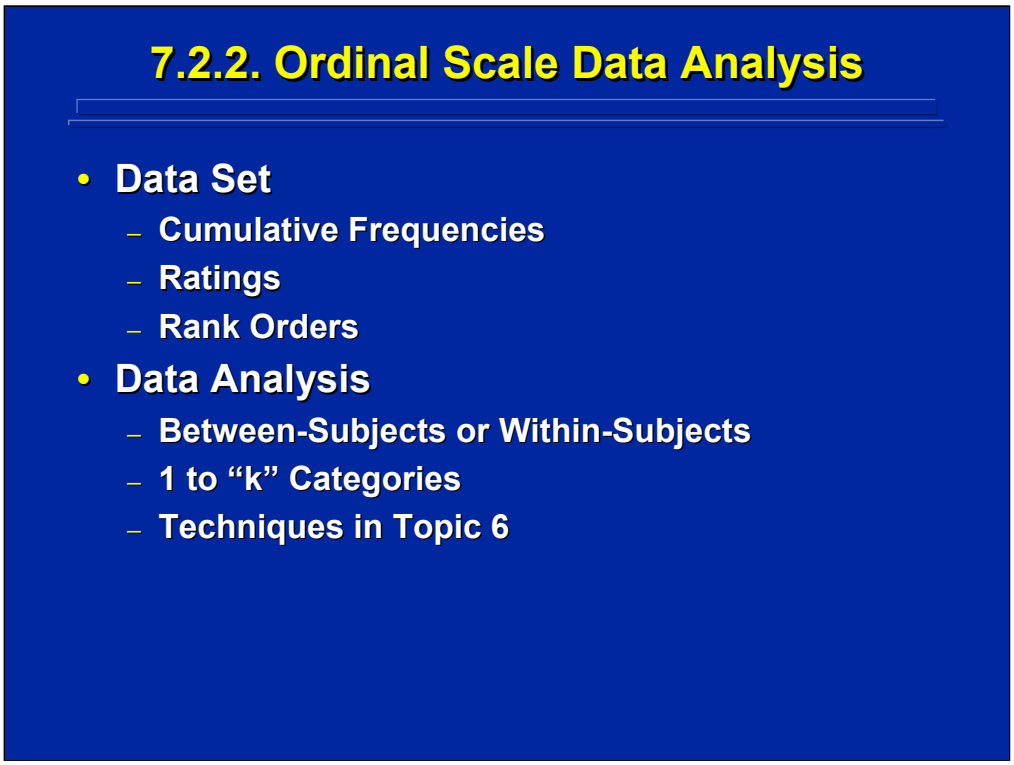



Topic 4 in this reference material provides an overview of the properties of various scales of measurement. Since supplemental data usually does not have properties of interval and ratio scale measurements, parametric data analyses are not appropriate. Consequently, human factors and ergonomic researchers usually use various nonparametric statistical analyses for supplemental data characterized by either nominal or ordinal scales of measurement.



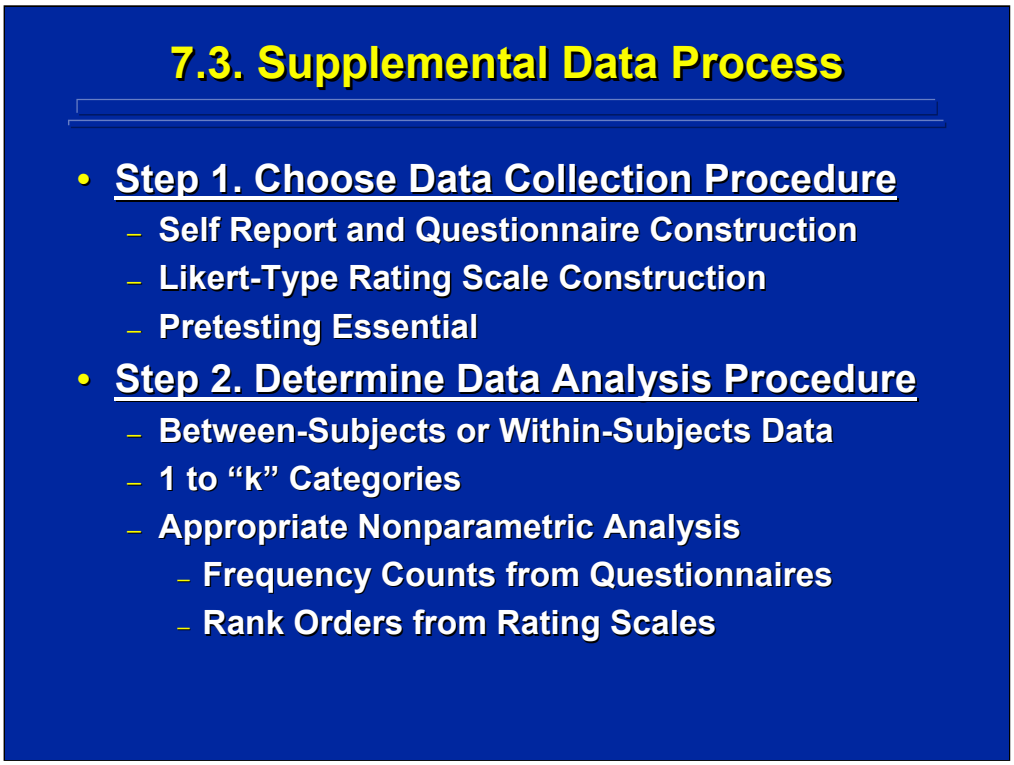

Topic 5 in this reference material describes a variety of nonparametric analysis procedures appropriate for nominal scale measurement. Nominal scale supplemental data is characterized by frequency counts within categories resulting from demographic data such as age and level of experience of users or frequency counts of various questionnaire alternatives.

The choice of nonparametric analysis alternative depends upon whether the data set consists of between-subjects or within-subject data and the number of categories being compared. Topic 5 classifies and describes some of common nonparametric analyses for nominal scale supplemental data collected in human factors research by type of data set and number of categories to facilitate easy reference by users of this reference material.



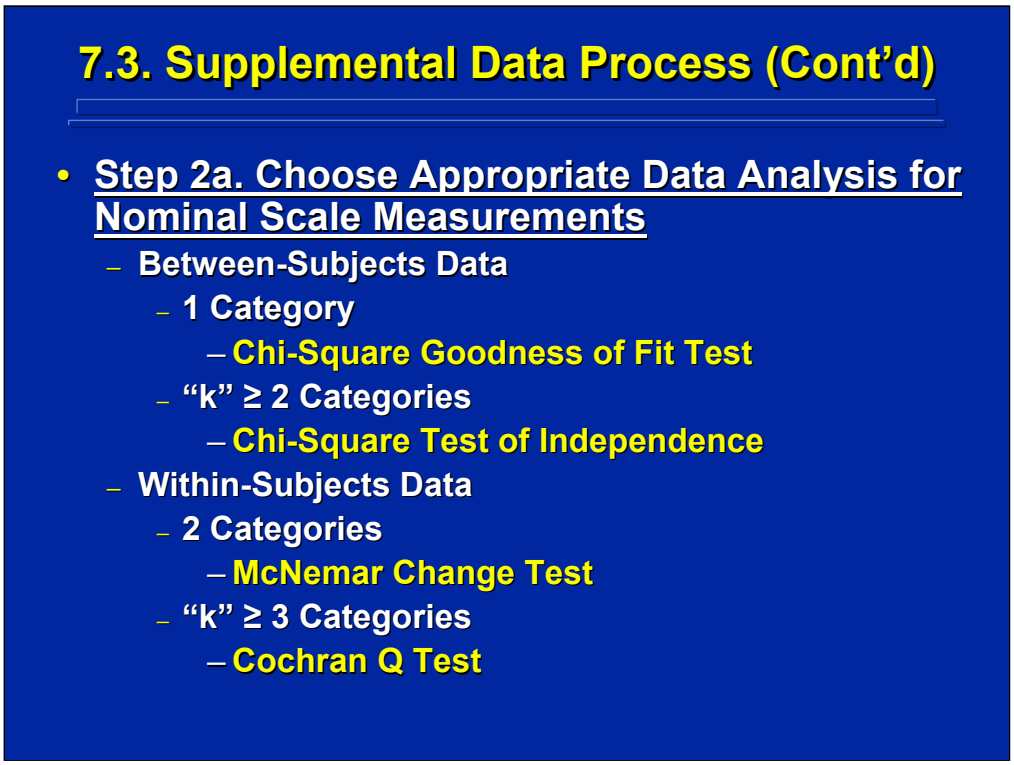

Topic 6 in this reference material describes a variety of nonparametric analysis procedures appropriate for ordinal scale measurement. Ordinal scale supplemental data is characterized by frequency counts within categories resulting from demographic data such as cumulative frequencies of ordered categories, graphical numerical rating scales, and rank orders.

The choice of nonparametric analysis alternative depends upon whether the data set consists of between-subjects or within-subject data and the number of categories being compared. Topic 6 classifies and describes some of common nonparametric analyses of ordinal scale supplemental data collected in human factors research by the type of data set and the number of categories to facilitate easy reference by users of this reference material.




The experimenter needs to consider an overall process in choosing the data collection and data analysis procedure for dealing with supplemental data. A three-step process is presented on this slide and the next two slides. This process begins with choosing the appropriate data collection procedure in Step 1. In order to provide quantitative data for subsequent data analysis, structured self report techniques, closed-ended questionnaires, and/or Likert-type rating scales as described in Topic 4 are often used in human factors and ergonomics research. Careful design and pretesting these data collection techniques are essential in order to obtain valid and reliable supplemental data.

Once the supplemental data are collected, the experimenter chooses the appropriate data analysis procedure in Step 2. The choice of the appropriate data analysis depends on three conditions – between-subjects or within-subjects data, number of categories evaluated, and scale of measurement as described in Topic 4. Since supplemental data usually includes only nominal or ordinal scales of measurement, nonparametric statistical analyses are used. Choosing the appropriate nonparametric analysis depends upon the use nominal and ordinal scale measurements as detailed in Steps 2a and 2b in the next two slides.



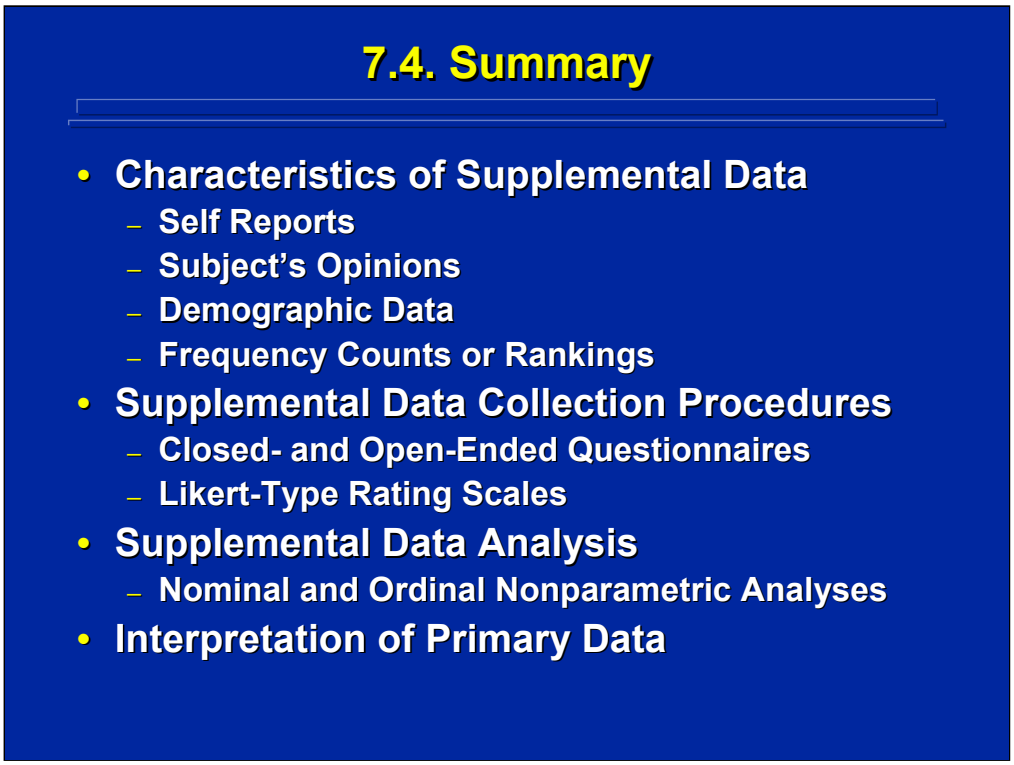

Step 2a provides a procedure for choosing the appropriate nonparametric analysis if the supplemental data involve nominal scale measurements. The choice depends upon the data set and number of categories compared. The Chi-Square Goodness of Fit Test is used to compare one category of data of various levels each obtained from independent samples of subjects (i.e., between-subjects data) to a known standard. The Chi-Square Test of Independence is used to compare two or more categories each with various levels obtained from independent samples. The McNemar Change Test uses repeated measures (i.e., within-subjects data) to evaluate differences or before/after changes between two categories. The Cochran Q Test extends the McNemar Change Test to within-subjects data that involve three or more categories. Details on calculations and examples of using each of these four nonparametric tests listed on this slide are provided in Topic 5.

## 7.3. Supplemental Data Process (Cont'd)

- **Step 2b. Choose Appropriate Data Analysis for Ordinal Scale Measurements**
  - Between-Subjects Data
    - 1 or 2 Categories
      - Kolmogorov-Smirnov Tests
    - "k" ≥ 3 Categories
      - Kruskal-Wallis One-Way ANOVA
  - Within-Subjects Data
    - 2 Categories
      - Wilcoxon Signed Ranks Test
    - "k" ≥ 3 Categories
      - Friedman Two-Way ANOVA
- **Step 3. Interpret Supplemental Data**

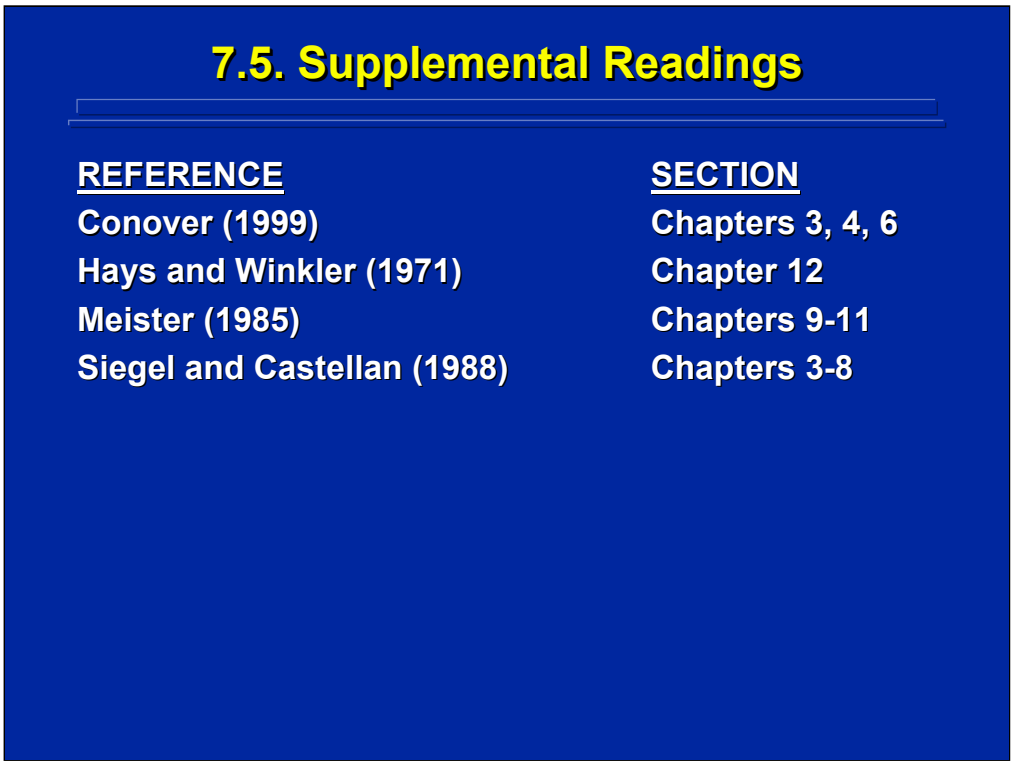
Step 2b provides a procedure for choosing the appropriate nonparametric analysis if the supplemental data involve ordinal scale measurements. The choice depends upon the data set and number of categories compared. Kolmogorov-Smirnov Tests of cumulative frequency distributions are used for between-subjects data to compare either one category of data to a standard or two categories of data. The Kruskal-Wallis One-Way ANOVA is used for analysis of three or more between-subjects categories. The Wilcoxon Signed Rank Test is used to evaluate differences between two categories of within-subjects data. The Friedman Two-Way ANOVA evaluates repeated measures data across three or more categories. Details on calculations and examples of using each of these four nonparametric tests listed on this slide are provided in Topic 6.

Step 3 deals with supplemental data interpretation. First, the results of the statistical analysis of the supplemental data need to be interpreted. Second, and most importantly, the supplemental data results need to be used by the experimenter to clarify the interpretations of the main analyses of the experiment. The appropriate main analyses for the overall experimental design chosen by the experimenter are described in the remaining sections of this reference material.




By way of summary, Section 2 deals with supplemental data collected in support of the primary data collected in the experiment. Supplemental data consist of subject's opinions and self reports, demographic data, frequency tabulations, and ratings. Experimenters should use carefully designed subjective data collection methods that provide quantitative data, if possible.

Usually a combination of self report methods including closed-ended and open-ended questionnaires, and Likert-type rating scales are used in human factors and ergonomics experiments. The experimenter must be careful to pretest these methods for clarity of wording and instruction before collecting the supplemental data. Most supplemental data consist of frequency, rating, and rankings that have only nominal or ordinal scale characteristics. Consequently, nonparametric rather than parametric statistical analyses are usually used to analyze these data. The experimenter should always remember the purpose of the supplemental data and use them to aid in the interpretation of the primary data collected in the experiment.




The Meister (1985) reference provides a general overview of various techniques used to collect supplemental data in human factors and ergonomics research. The Siegel and Castellan (1988) reference provides a detailed discussion of the nonparametric analysis techniques appropriate for nominal and ordinal scale data as described in Topics 5 and 6, respectively. Conover (1999) and Hays and Winkler (1971) are other general references on nonparametric analyses that provides further elaboration of the techniques covered in Section 4 of this reference material.

# Section 3.
# Basic Analysis of Variance (ANOVA) Designs

**Topic 8. Introduction to ANOVA**
**Topic 9. ANOVA Summary Table**
**Topic 10. Between-Subjects ANOVA Designs**
**Topic 11. Analysis of Comparisons and Interactions**
**Topic 12. Within-Subjects ANOVA Designs**
**Topic 13. Mixed-Factors ANOVA Designs**
**Topic 14. Summary of Basic ANOVA**

Section 3 covers fundamental experimental design and analysis procedures used in basic ANOVA. These designs are the most often used techniques in human factors and ergonomics research. This section covers the following topics:

Topic 8 – introduction to ANOVA;

Topic 9 – ANOVA summary table components;

Topic 10 – between-subjects ANOVA designs;

Topic 11 – analysis of comparisons and interactions;

Topic 12 – within-subjects ANOVA designs;

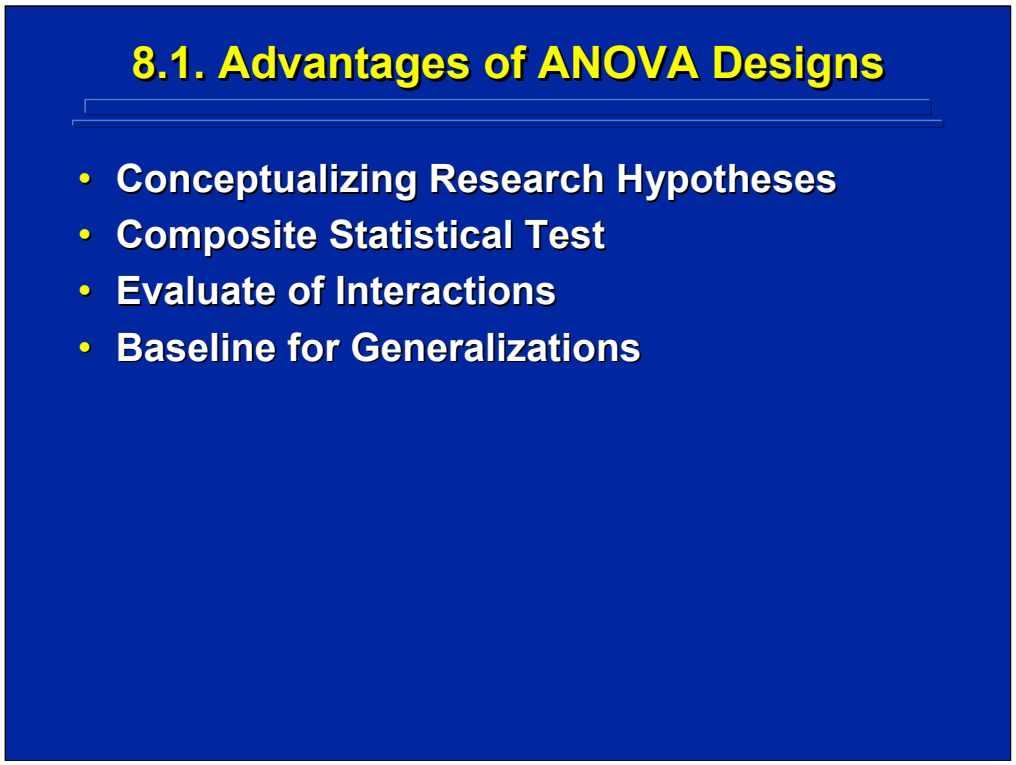
Topic 13 – mixed-factors ANOVA designs; and

Topic 14 – summary of basic ANOVA.

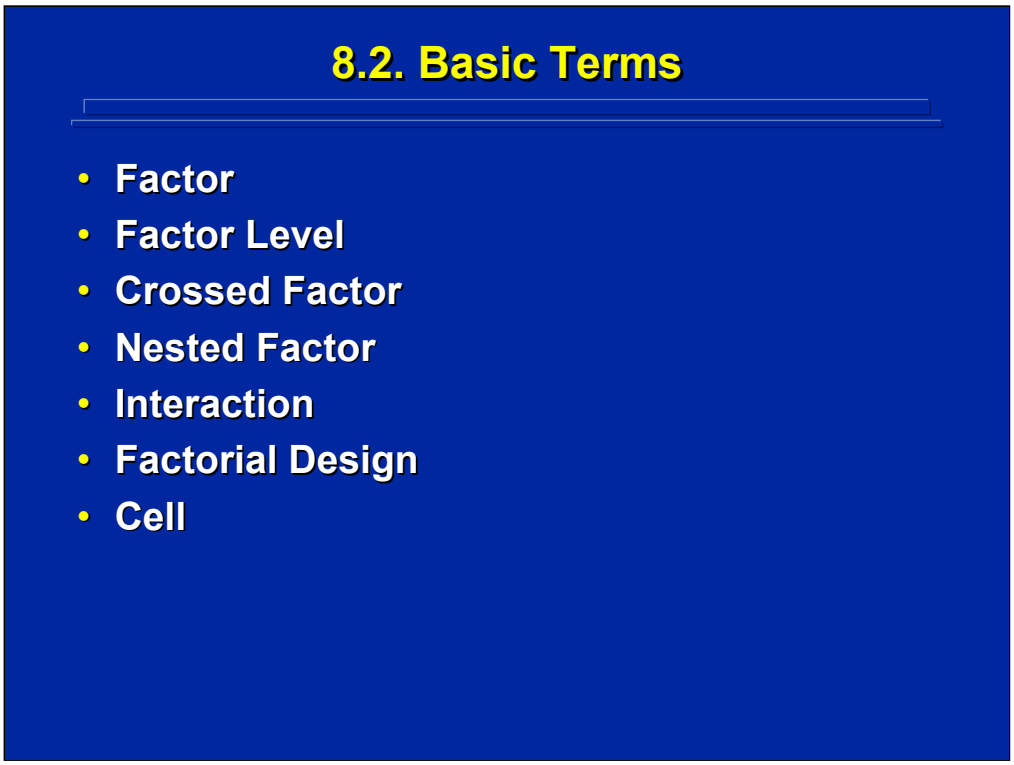
# Topic 8. Introduction to ANOVA

8.1. Advantages of ANOVA Designs
8.2. Basic Terms
8.3. ANOVA Design Alternatives
8.4. ANOVA Statistical Models
8.5. ANOVA Hypothesis Testing
8.6. Summary
8.7. Supplemental Readings

This topic introduces ANOVA designs by discussing their advantages, basic terms, and the three major categories of ANOVA designs used in human factors research. Next, procedures for specifying the underlying statistical model that describes the components of any ANOVA design are presented. The introduction ends with a discussion and example for using ANOVA for statistical hypothesis testing of the difference between two treatment means.

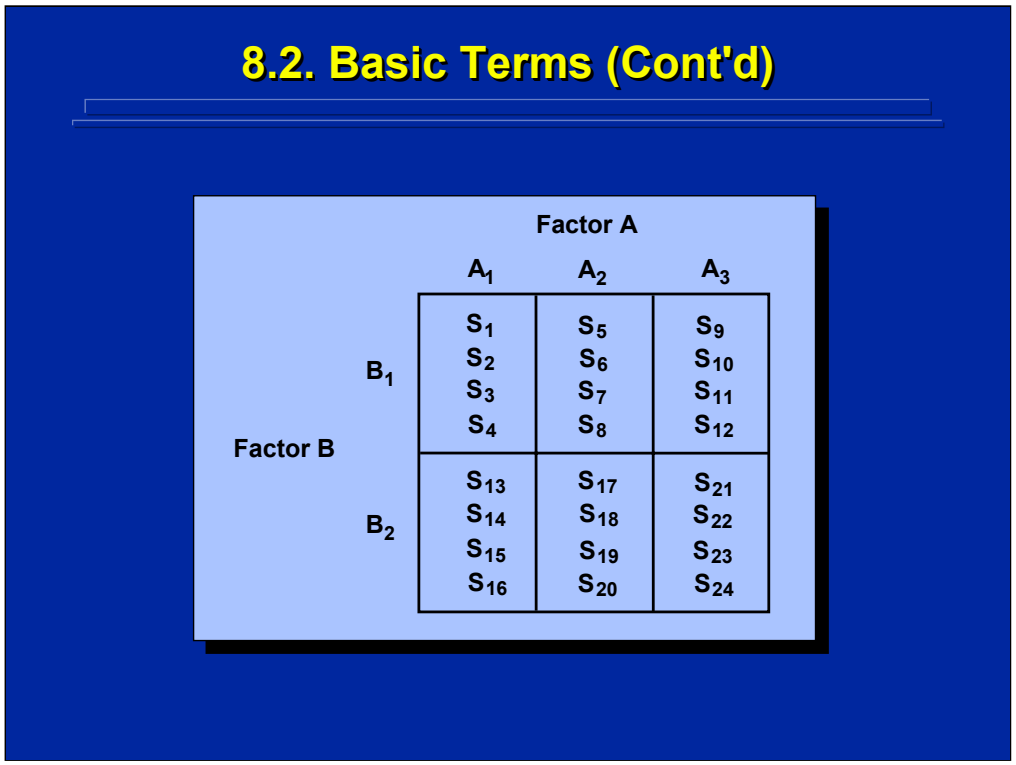



This slide lists the four major advantages of ANOVA. The researcher is forced to organize and conceptualize research hypotheses of interest when choosing the number of independent variables to include in the ANOVA design. As a means of guarding against inflated Type I error resulting from repeated hypothesis tests on the same data set, the ANOVA provides a composite test of significance of main effects and interactions of all the independent variables included in the design. The interaction, or differential effect, of one independent variable on others can be evaluated along with the main effects of each variable in a factorial ANOVA design. Since several independent variables can be investigated simultaneously in multifactor ANOVA designs, the researcher has a broader baseline for making generalization to real-world problems based on the results obtained from ANOVA designs. Because of these advantages, ANOVA is the most often used experimental design alternative in human factors and ergonomics research.




Williges (1995) defined seven key terms used to describe ANOVA designs. These terms form the basic vocabulary needed to specify and describe ANOVA designs used in human factors research. A factor is an independent variable manipulated in the design (e.g., display type). Subjects always appear as a factor along with other factors of interest in human factors experiments. A specific value of a factor is known as the factor level, and all factors must have a minimum of two levels (e.g., plasma and liquid crystal displays). Factors are crossed if all the levels of one factor appear with all the levels of another factor (e.g., every subject receives every treatment). Factors are nested if only one level of a factor appears at each level of another factor (e.g., each subject receives only one treatment). An interaction is a differential effect of one factor on another such that the levels of one factor are significantly different only at a particular level of the other factor. Factors must be crossed in order to interact. If factors are nested, no interaction can be evaluated.

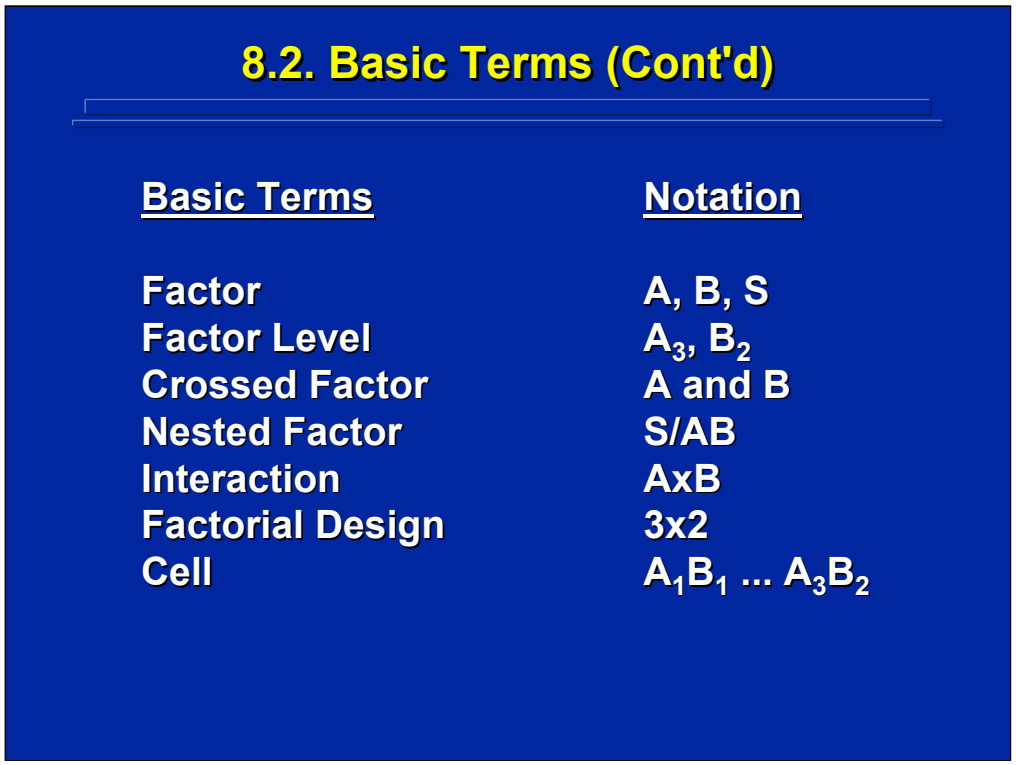
A factorial design is a design in which all the levels of one factor appear with all the levels of another factor. Hence, the factors of interest are crossed. A unique treatment combination of a specific value of the various levels of factors in a factorial design is referred to as a cell of the design. Factorial designs are used in human factors research to assess interactions among the factors of interest in addition to the main effects of these factors.



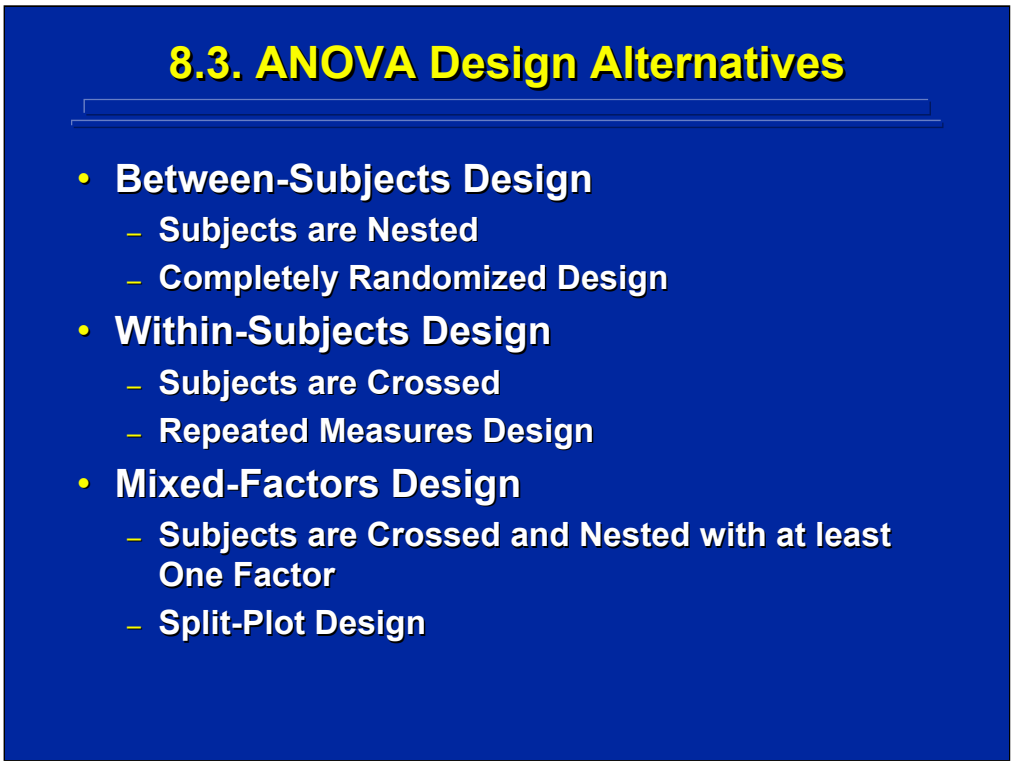

This is an illustration of a 3x2 factorial design that can be described by using the basic ANOVA terms. By convention, factors are designated as capital letters, and levels of a factor are designated as subscripted numbers of the capital letter. There are three levels of factor A and two levels of factor B. Since factors A and B are crossed, each level of factor A appears at each level of factor B resulting in six cells or treatment combinations in the complete factorial design. An equal cell size, n, of four is used.

A third factor implicit in this design is subjects, S. There are 24 levels of factor S shown in the design. Subjects are nested, not crossed, with factors A and B since only four different subjects appear in each of the six cells of the factorial design. Note that each subject experiences only one combination of levels of A and B, but all the levels of A and B are crossed in the factorial design. Consequently, this is a 3x2 factorial design in which subjects are nested in both factors A and B.

Since subjects will always be a factor in human factors experiments, one must know whether subjects are crossed or nested with each of the factors of interest. This determines the sample size, n, in each cell of the design as well as the total number of different subjects needed for the experiment. In this 3x2 factorial design, four different subjects experienced each combination of Factor A and B levels yielding a total of 24 different subjects participated in the experiment.




This slide illustrates how the seven basic ANOVA terms are used in the 3x2 factorial design example shown on the previous slide. Factors of interest are listed with the capital letters, A and B, and the subject factor is listed as S. Factor A has three levels and Factor B has 2 levels. Specific levels of a factor appear as numbered subscripts of capital letters for that factor (e.g., $A_3$). Both A and B are crossed factors. Subjects factor, S, is nested in both A and B and is designated by a slash, S/AB. Only one interaction, AxB, can be tested in this design. Note that the other possible two-way interactions (i.e., AxS and BxS) and the three-way interaction (i.e., AxBxS) do not exist in this design because subjects are not crossed with either A or B. The factorial design is designated by the number of levels of the crossed factors of interest, 3x2. A cell of the factorial design is designated by each of the six unique treatment combinations of the various levels of Factor A and B or $A_1B_1$ … $A_3B_2$.




How subjects are crossed and nested with the factors of interest in an experiment determines the design category. There are three basic design alternatives in behavioral research. If subjects are nested within all factors of interest in the experiment, this is a between-subjects design or a completely randomized design because subjects are randomly assigned to treatment conditions. If subjects are crossed with all factors of interest, this is referred to as a within-subjects design or repeated measures design because every subject experiences every treatment combination. If subjects are nested within some factors of interest and crossed with others, this is called a mixed-factors design or split-plot design from agricultural applications where factors were split within plots of land.

Often the experimenter can choose a between-subjects, within-subjects, or mixed-factors ANOVA design and must then trade off the advantages and disadvantages of each design alternative. Sometimes factors exist only as crossed with subjects (e.g., practice trials) or nested with subjects (e.g., training method) in the real world and no choice of design is possible. Due to the nature of variables investigated in human factors research, the experimenter often chooses a mixed-factors ANOVA design alternative.

## 8.3. ANOVA Design Alternatives (Cont'd)

- **Between-Subjects Design**

| Factor B | | Factor A: $A_1$ | Factor A: $A_2$ | Factor A: $A_3$ |
|---|---|---|---|---|
| | $B_1$ | $S_1$ $S_2$ $S_3$ $S_4$ | $S_5$ $S_6$ $S_7$ $S_8$ | $S_9$ $S_{10}$ $S_{11}$ $S_{12}$ |
| | $B_2$ | $S_{13}$ $S_{14}$ $S_{15}$ $S_{16}$ | $S_{17}$ $S_{18}$ $S_{19}$ $S_{20}$ | $S_{21}$ $S_{22}$ $S_{23}$ $S_{24}$ |

Returning to the 3x2 factorial design presented when discussing basic terms of ANOVA in 7.2, one knows this is a between-subjects design because the subjects are nested in both A and B. Four subjects (i.e. n = 4) receive each of the six treatment combinations in this factorial design. Since each subject experiences only one treatment combination in a between-subjects design, a total of 24 different subjects are needed for this experiment.

## 8.3. ANOVA Design Alternatives (Cont'd)

- **Within-Subjects Design**

| | | Factor A | | |
|---|---|---|---|---|
| | | $A_1$ | $A_2$ | $A_3$ |
| Factor B | $B_1$ | $S_1$ $S_2$ $S_3$ $S_4$ | $S_1$ $S_2$ $S_3$ $S_4$ | $S_1$ $S_2$ $S_3$ $S_4$ |
| | $B_2$ | $S_1$ $S_2$ $S_3$ $S_4$ | $S_1$ $S_2$ $S_3$ $S_4$ | $S_1$ $S_2$ $S_3$ $S_4$ |

This slide shows the 3x2 factorial design cast as a within-subjects design with n=4. One knows this is a within-subjects design because the subjects are crossed with both A and B. Consequently, each subject appears in every cell, and only four different subjects are needed for the experiment. When using a within-subjects design, the experimenter must balance the order in which each subject receives the six treatment combinations to avoid confounding practice effects with the treatment conditions.

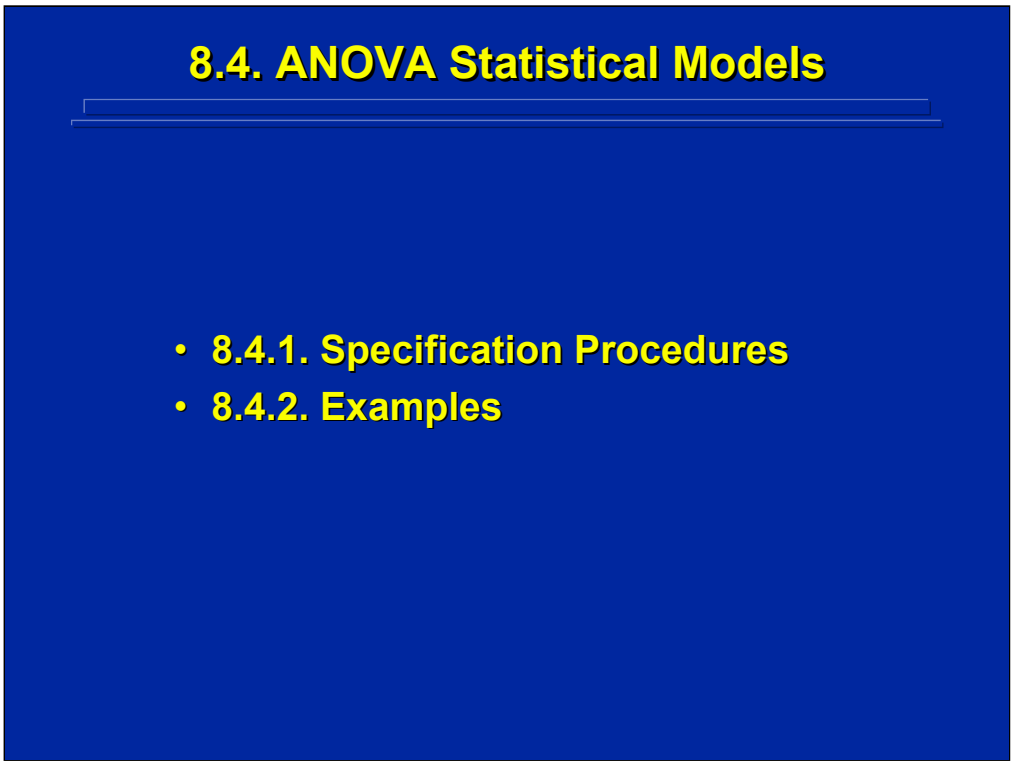
## 8.3. ANOVA Design Alternatives (Cont'd)

- **Mixed-Factors Design**

| Factor B | Factor A: $A_1$ | Factor A: $A_2$ | Factor A: $A_3$ |
|---|---|---|---|
| $B_1$ | $S_1$ $S_2$ $S_3$ $S_4$ | $S_5$ $S_6$ $S_7$ $S_8$ | $S_9$ $S_{10}$ $S_{11}$ $S_{12}$ |
| $B_2$ | $S_1$ $S_2$ $S_3$ $S_4$ | $S_5$ $S_6$ $S_7$ $S_8$ | $S_9$ $S_{10}$ $S_{11}$ $S_{12}$ |

This slide shows the 3x2 factorial design cast as a mixed-factors design with n=4. By looking at the subscripts of S in each cell, one can determine that subjects are nested within Factor A and crossed with Factor B. Each subject experiences only one level of Factor A but both levels of Factor B. Consequently, each subject receives two treatment combinations (i.e., a level of Factor A with each of the two levels of Factor B). A total of 12 different subjects are needed for the experiment.

If the 3x2 mixed-factors design were changed such that subjects were crossed with Factor A and nested in Factor B, the subscripts of S would change accordingly. Four subjects (i.e., $S_1$ … $S_4$) in $B_1$ and four different subjects (i.e., $S_5$ … $S_8$) in $B_2$ would each receive the three levels of Factor A in combination with only one level of Factor B. Consequently, a total of eight different subjects would be needed for this experiment. Once again, the experimenter needs to balance the presentation order of the within-subjects factor levels to avoid confounding practice effects with treatments.




Every ANOVA design can be specified in terms of a statistical model that defines the various components that can affect an observation, Y, in the experimental design. This subsection describes the procedures for specifying ANOVA statistical models and provides an example of statistical models for two-factor between-subjects, within-subjects, and mixed-factors ANOVA designs.

## 8.4. ANOVA Statistical Models (Cont'd)

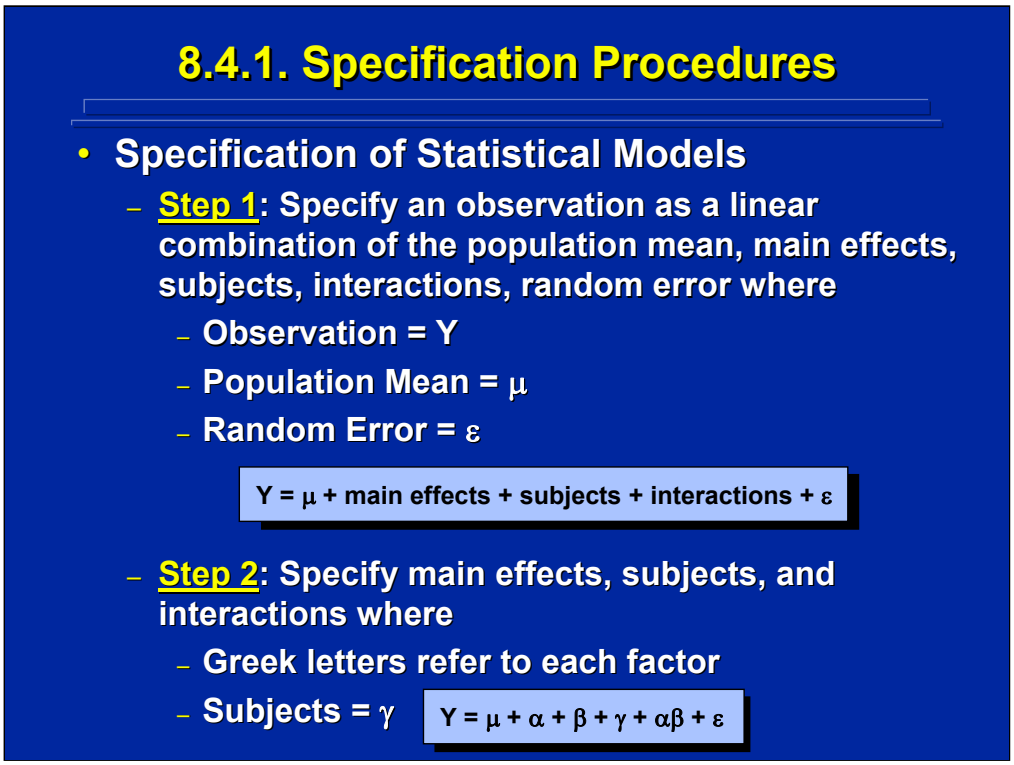
- **Definition: A mathematical statement expressing the linear sum of all possible components of variation in a specific experiment.**

  $Y = \mu + \text{Main Effects} + \text{Subjects} + \text{Interactions} + \varepsilon$

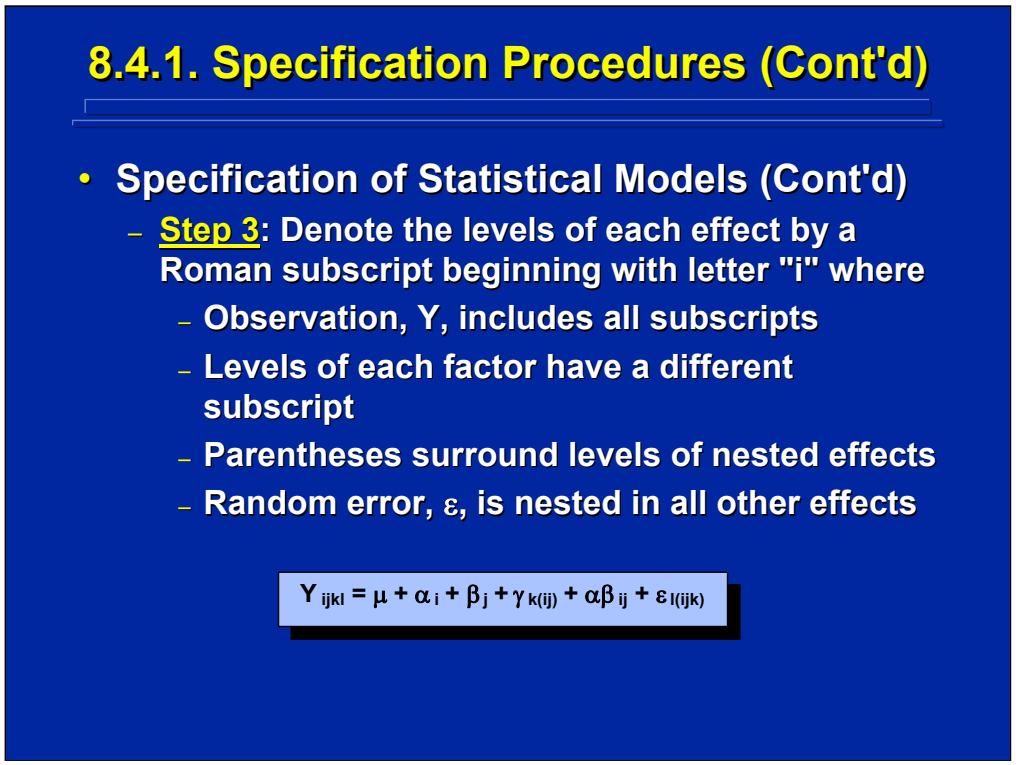
- **Components**
  - **Observation, Y**
  - **Population Mean, $\mu$**
  - **Factor Main Effects**
  - **Subject Effect**
  - **Interaction Effects**
  - **Random Error, $\varepsilon$**

The ANOVA statistical model is a mathematical statement that lists all the possible components of variation in a specific experiment (Keppel and Wickens, 2004 and Montgomery, 2005). Winer, Brown and Michels (1991) refer to statistical models as structural models. The resulting statistical model is simply a linear sum or combination of sources of variation that can affect any observed score, Y, obtained from subjects in the experiment. The major components of an observation in a human factors experiment are: (1) the population mean from which the sample is drawn; (2) the factors and interactions of interest to the experiment; (3) the subject effects; and (4) random error. All subsequent analyses using deviation scores are based on the statistical model. Before conducting an experiment, the experimenter should specify the ANOVA statistical model to insure that all factors and interactions of interest are included in the experimental design.

An alternative approach to partitioning variation through the ANOVA statistical model is the general linear model based on regression. Keppel and Wickens (2004, pp. 132-158) describe the use of general linear models in ANOVA that are used in many computer-based statistical analysis procedures. Tatsuoka (1993, pp. 3-42) compares the general linear model to the variance component models of ANOVA. Statistical models of variance components facilitate computational procedures based on equal sample size and are used in this reference material to describe general models of various ANOVA designs.

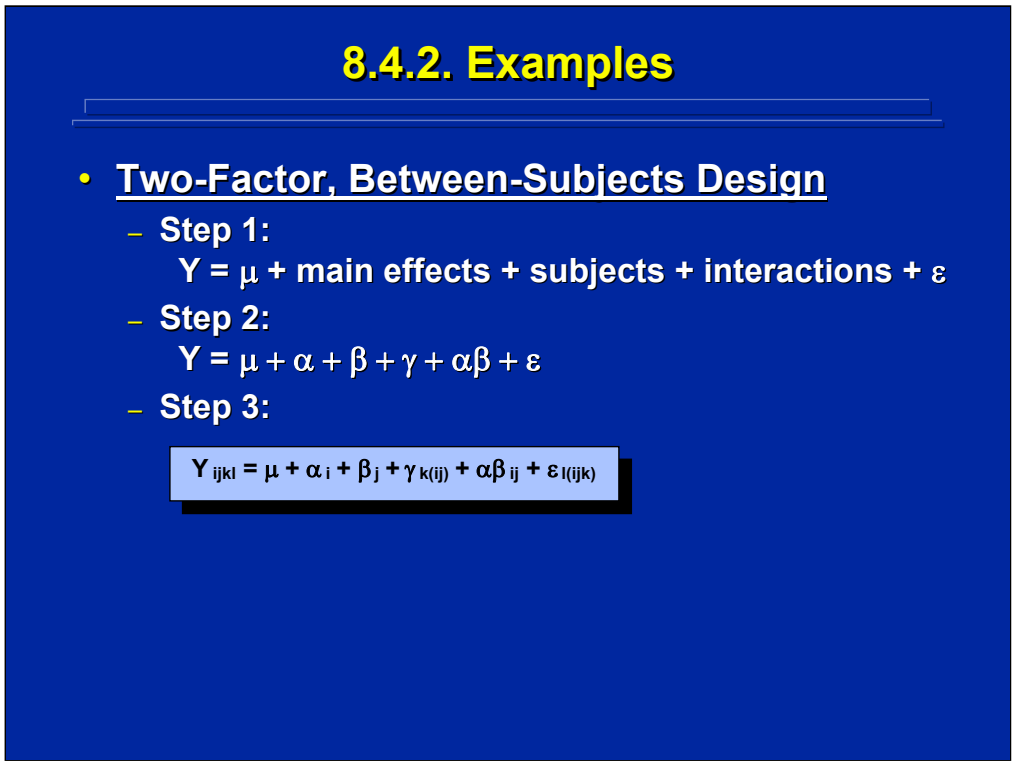



There are straightforward procedures for specifying ANOVA statistical models. One specifies an observation, Y, as being equal to a linear combination of the population mean, main effects, interactions, and random error. Greek letters are used to define each component. To simplify reading the statistical model, it begins with the population mean, $\mu$, and ends with random error, $\varepsilon$. All the factor main effects, subjects, and possible interactions are listed between $\mu$ and $\varepsilon$. Greek letters beginning with alpha, $\alpha$, are used to specify each factor and gamma, $\gamma$, is reserved to specify the subject effect in the experiment. The equation shown at the bottom of this slide in Step 2 represents a two-factor, between-subjects design where only Factors A and B can interact.




Lowercase Roman letters beginning with “i” are used to denote specific levels of each component. The observation, Y, includes all subscripts. Each factor is denoted with a different subscript, and the nesting among factors is designated by parentheses. Usually factors of interest are crossed in factorial designs, and only subjects and random error show nesting. Random error is always nested within all other effects by definition. Consequently, the subscripts for all effects are put in parentheses for the subscript designating random error. For example, the resulting statistical model for a two-factor, between-subjects design is shown at the bottom of this slide. Notice that the subscripts “ij” are put in parentheses for subjects, $\gamma$, to designate that subjects are nested in Factors A and B and cannot interact with those factors.

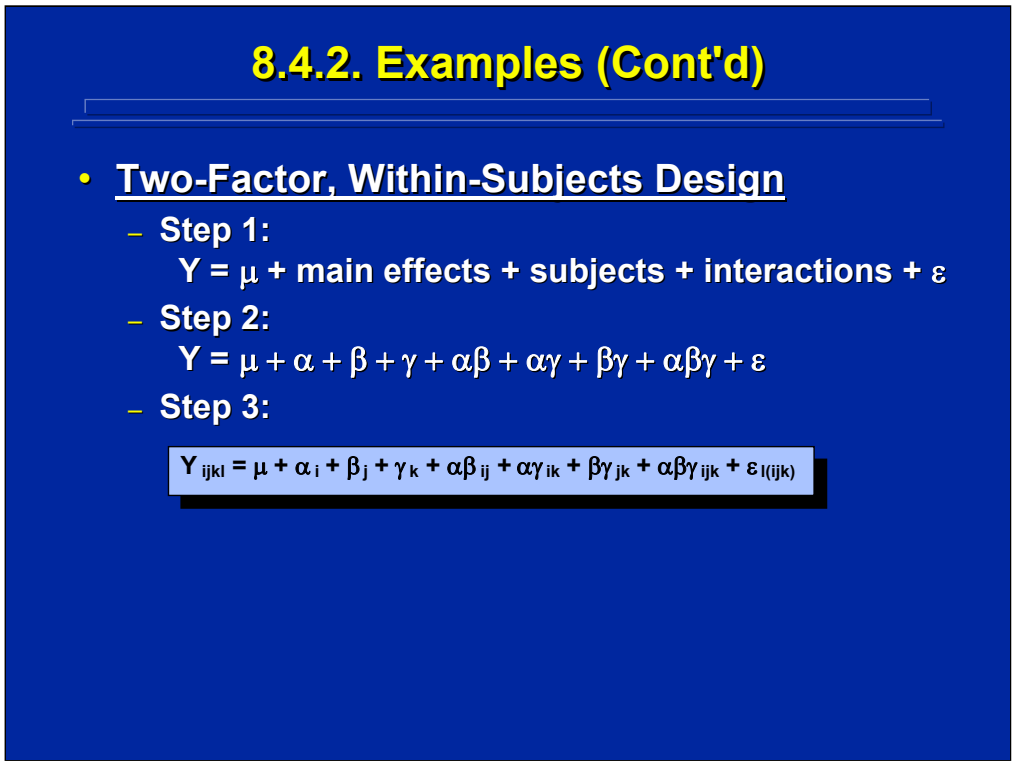
The key to determining whether the ANOVA design is a between-subjects, within-subjects, or mixed-factors design is to designate the nesting of subjects, $\gamma$, appropriately. To illustrate this concept, two-factor ANOVA design examples are provided for each of these three categories of the experimental design separately.




The two-factor between-subjects design is shown on this slide. Since subjects are nested within all other effects, there can be no interaction of S with A, B, or random error. Keppel and Wickens (2004) and Montgomery (2005) do not list subject effects in their statistical models of randomized designs, but the nested subject effect is always listed in the statistical models in this reference material to clearly distinguish between-subjects designs. Only A and B can interact because they are crossed in this design. Consequently, the statistical model for this between-subjects design contains the population mean ($\mu$), the main effect of A ($\alpha$), the main effect of B ($\beta$), the main effect of subjects ($\gamma$), the AxB interaction ($\alpha\beta$), and random error ($\varepsilon$) as shown in Step 2.

The final step is to add subscripts to the statistical models to designate crossed and nested effects in the design. The observation, Y, is influenced by all effects in the statistical model and includes the subscripts “ijkl”. Factor A begins with the “i” subscript, and Factor B continues with the “j” subscript. Both A and B have no nesting. For S the subscript is “k”, but since it is nested in both A and B, this nesting is designated by placing ij in parentheses, or k(ij). The AxB interaction represents both A and B effects and has the subscript “ij”. Random error is nesting in all effects, and its subscript is l(ijk).

The final statistical model for this design is shown in Step 3. Note the subscripts for gamma show this is a between-subjects design because S is nested in both Factors A and B.




The two-factor, within-subjects design is shown on this slide. The same three steps are followed for specifying the ANOVA statistical model. Since A, B, and S are completely crossed in this repeated measures design, no nesting is designated in the subscripts. Note that S just has the subscript “k” without any parentheses to show that it is not nested. Consequently, the statistical model of the within-subjects version of a two-factor design must also include the three possible interactions with subjects (i.e. $\alpha\gamma_{ik}$ for AxS, $\beta\gamma_{jk}$ for BxS, and $\alpha\beta\gamma_{ijk}$ for AxBxS).

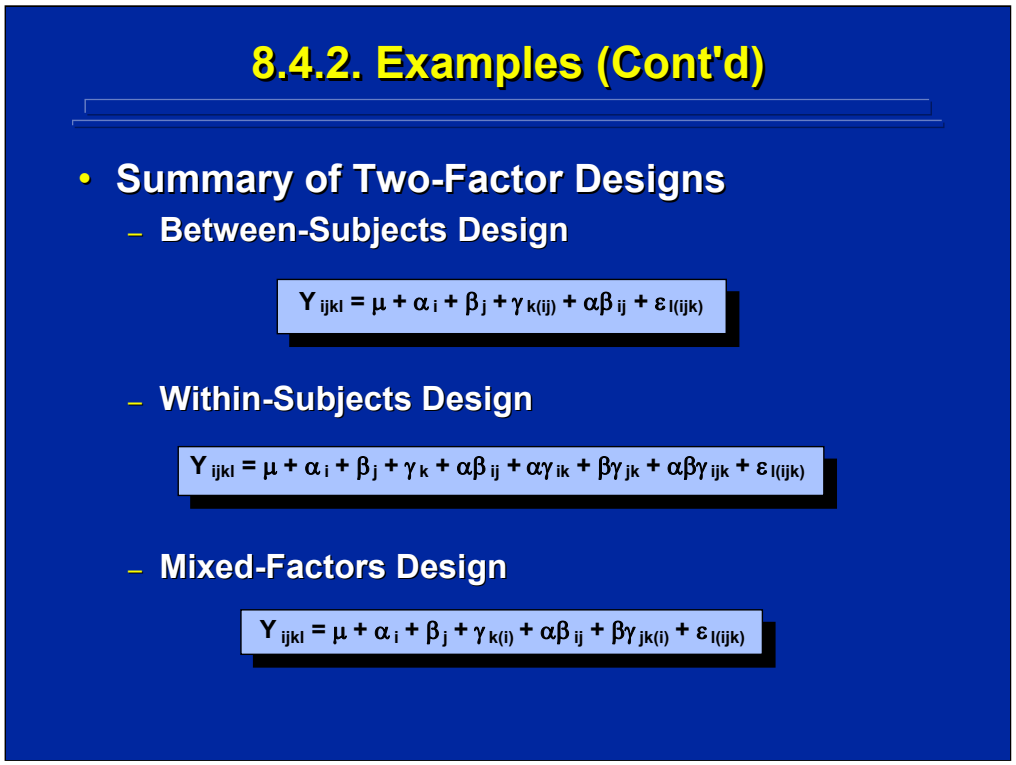
## 8.4.2. Examples (Cont'd)

- **Two-Factor, Mixed-Factors Design where A = Between-Subjects, B = Within-Subjects**
  - **Step 1:**
    $Y = \mu + \text{main effects} + \text{subjects} + \text{interactions} + \varepsilon$
  - **Step 2:**
    $Y = \mu + \alpha + \beta + \gamma + \alpha\beta + \beta\gamma + \varepsilon$
  - **Step 3:**

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_{k(i)} + \alpha\beta_{ij} + \beta\gamma_{jk(i)} + \varepsilon_{l(ijk)}$$

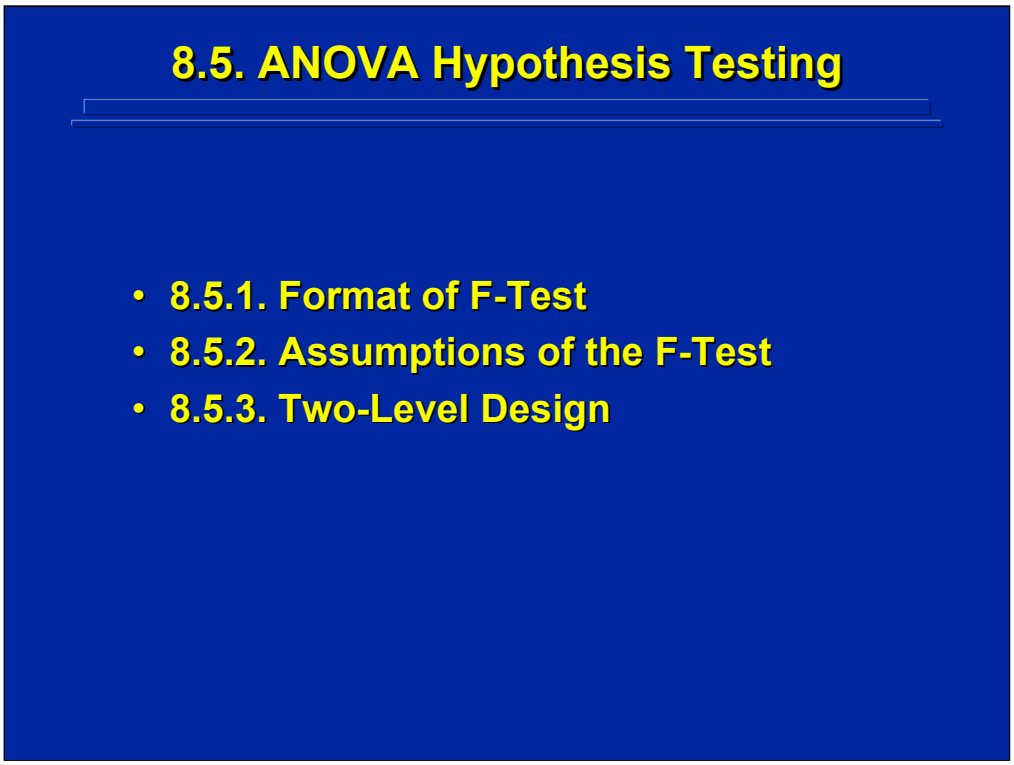
A two-factor, mixed-factors design is shown on this slide. Factor A is a between-subjects variable, and B is a within-subjects variable. Consequently, $\gamma$, or S, shows this nesting by using the subscript "k(i)". In this two-factor ANOVA design, S cannot interact with A. So, no $\alpha\gamma$ effect is shown in the statistical model.

If this two-factor design was reversed to make A a within-subjects factor and B a between-subjects factor, the resulting ANOVA statistical model would change accordingly. Gamma would have the subscript "k(j)", $\beta\gamma$ would be replaced by $\alpha\gamma$ with the subscripts "ik(j)".

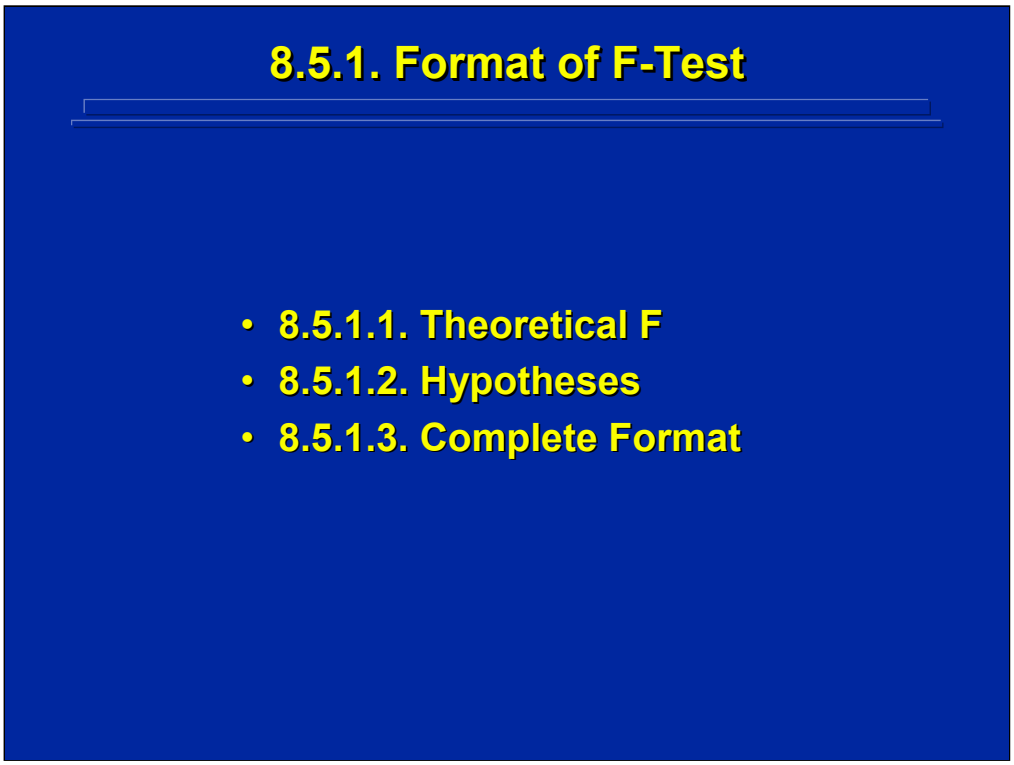



This slide summarizes the statistical models for the three alternative two-factor ANOVA designs. Note that the designs differ in the number of effects between the population mean and random error that can be estimated in the subsequent ANOVA. The final statistical model for $Y_{ijkl}$ shows that there are four effects that can be estimated in the between-subjects design, seven effects that can be estimated in the within-subjects design, and five effects that can be estimated in the mixed-factors design.

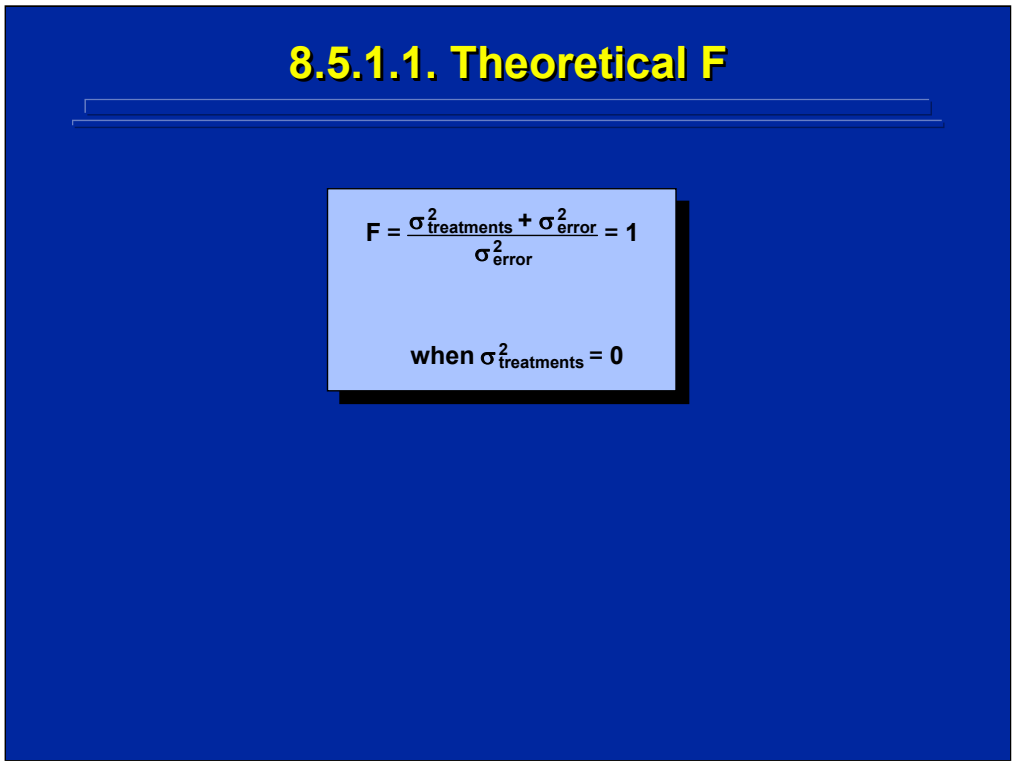
The key to specifying the ANOVA statistical model in human factors research is to determine how subjects will be assigned to treatment conditions. These crossed and nested relationships of subjects with the factors of interest in the experiment determine the number of effects in the resulting statistical model as well as the subscripting designations. In addition, the relationships dictate the number of different subjects required in an experiment to run an analysis based on an equal sample size of some value of n.




Before discussing the details of conducting hypothesis tests on the various components of an ANOVA statistical model, this subsection summarizes the basic format and assumptions of any ANOVA hypothesis test based on the F sampling distribution. The general logic followed in conducting a F-test is demonstrated in a simple, one-factor, two-level design.



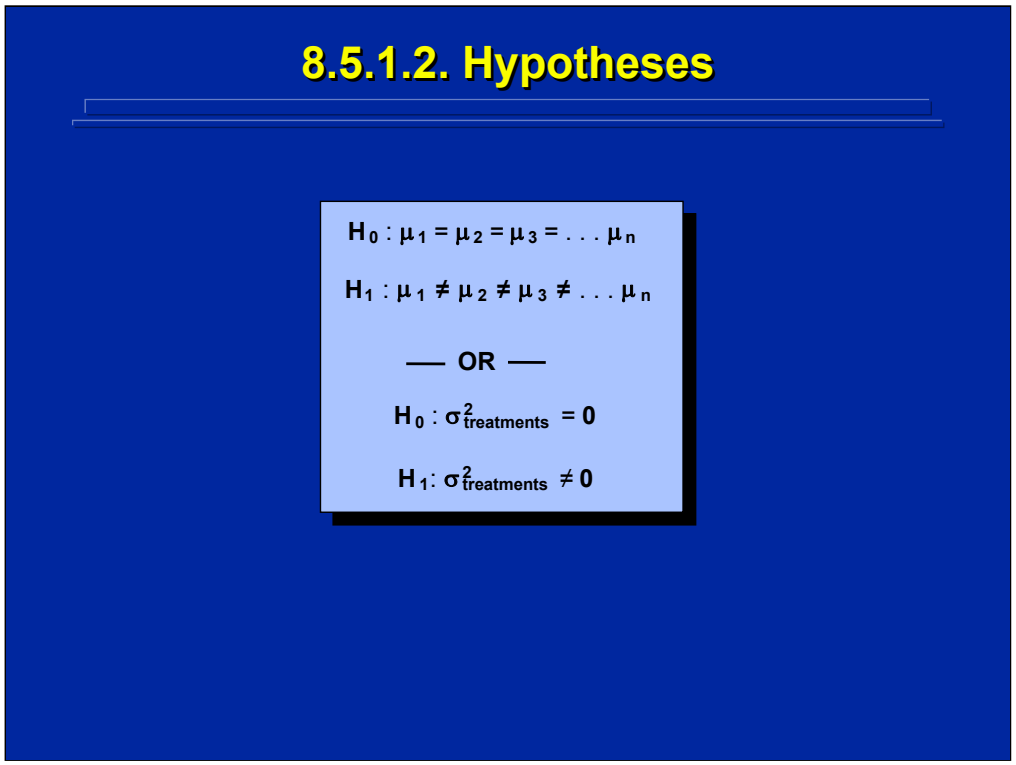

The F-test is a statistical hypothesis test using the F sampling distribution and the F statistic. The theoretical F value under the null hypothesis and the standard format for statistical hypothesis testing in ANOVA is described in this subsection.



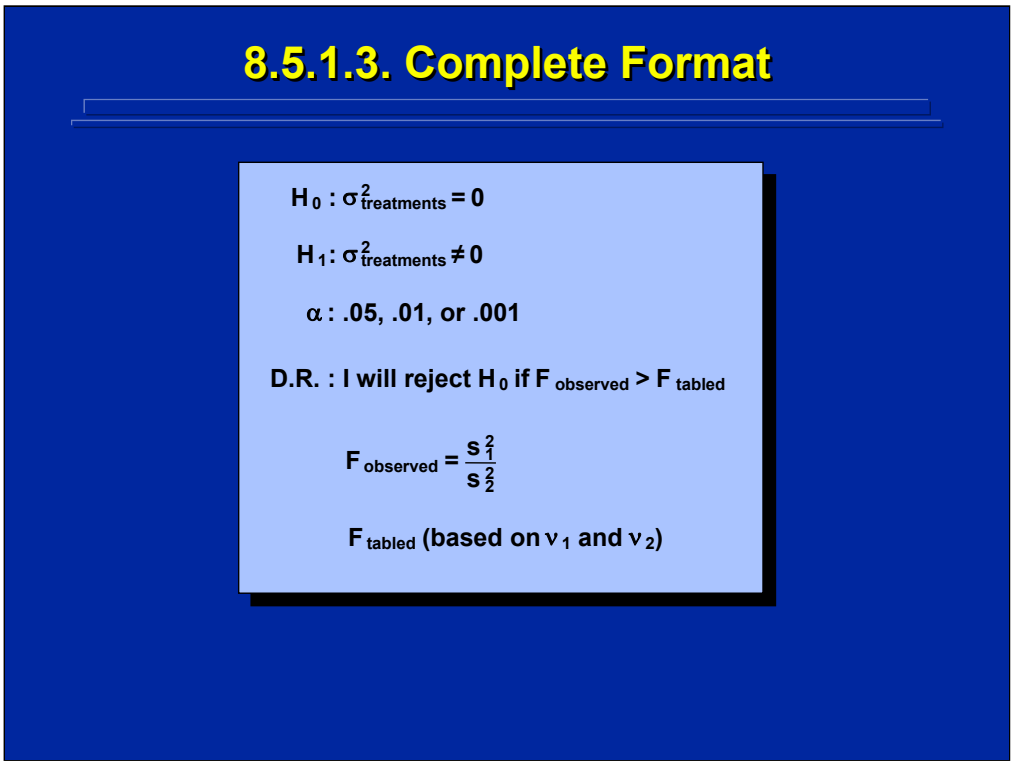

The F-statistic used in ANOVA is the ratio of two sample variances. One must decide which variance is used in the numerator and denominator of the F-ratio. As shown in this slide, theoretically the sample variance affected by the treatment effect and error is used in the numerator, and a sample variance affected only by error is used in the denominator. If a treatment effect does not exist, the variance due to treatments is 0.

In a statistical hypothesis test in ANOVA, one assumes under the null hypothesis that the treatment effect in the numerator does not exist. Consequently, the theoretical F value reduces to a ratio of two estimates of error variance and F equals 1. Occasionally, it is possible to obtain empirically an F-ratio that is less than 1 when the treatment does not exist and variation in random error is such that the estimate in the numerator is smaller than the denominator. If there is a treatment effect in the numerator, then the F value will be greater than 1.

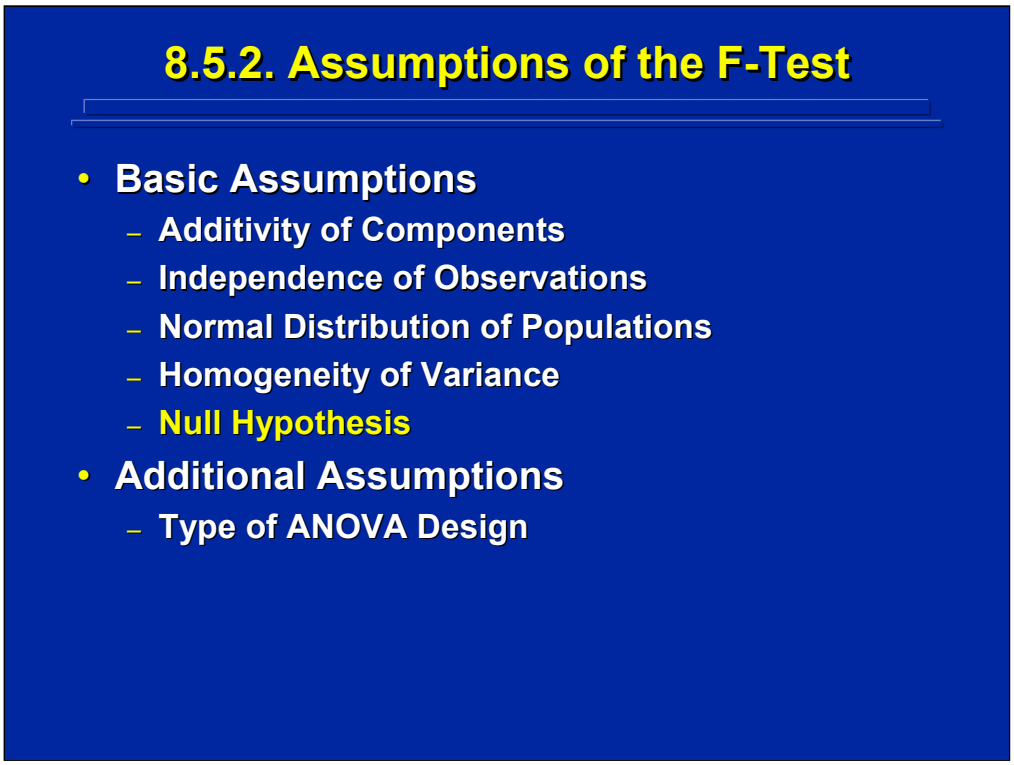
This is the basic premise of any hypothesis test using an F-statistic. Consequently, a researcher must determine which estimate of sample variance is put in the numerator and which estimate is put in the denominator in order to conduct a hypothesis test in ANOVA. Commonly, this is referred to as choosing the appropriate error term for an F-test.




In ANOVA, there are two or more levels of each factor, and the experimenter usually makes a composite test of differences across several means in a hypothesis test. The top portion of this slide shows the null and alternative hypotheses for a test among several means. The differences across means of factor levels determine the treatment effect of that factor. In ANOVA, one could alternatively state the null and alternative hypothesis in terms of the variance due to treatments as shown in the lower portion of this slide. This variance form is most commonly used in ANOVA, and the specific factor (i.e., A, B) or an interaction (i.e. AxB) is substituted for treatments.




The standard format used for an F-test is shown on the slide. As in any statistical hypothesis test, this format includes the null hypothesis ($H_0$), the alternative hypothesis ($H_1$), the amount of Type I error ($\alpha$) that one is willing to accept, and the decision rule. The decision rule in ANOVA can be simply stated that one will reject the null hypothesis if the observed F-statistic is greater than the tabled value of F. The observed F-statistic is calculated from the data collected in the experiment using the sample variance form of F. The tabled value of F is determined by the F sampling distribution based on the degrees of freedom of the numerator ($\nu_1$) and the denominator ($\nu_2$) of the F-ratio.




The basic assumptions of the F-test are shown on the top portion of this slide. The additivity assumption relates to the statistical model which states that an observation is based on several additive parts consisting of the population mean, treatment effects, subject effects, and random error. The assumptions of independence of observations, normal distribution of scores, and homogeneity of variance are based on the definition of an F statistic as the ratio of two independent chi-squares having equal population variance. The null hypothesis is that the variance due to treatments is equal to zero and is the basis of the F-test itself.

There are some additional assumptions that depend upon the specific type of ANOVA design used. For example, within-subjects designs assume homogeneity of covariance, and quasi-F tests have an additivity assumption. These additional assumptions are discussed under the topics covering these specific design alternatives.

## 8.5.2. Assumptions of the F-Test (Cont'd)

- Violation of Assumptions
  - Robustness of Analysis of Variance
  - F < 1 throughout ANOVA
  - Data Transformations for Normality
  - Homogeneity of Variance Tests
- Alternative Tests for Homogeneity of Variance
  - Hartley F-Max Test
  - Cochran Test
  - Bartlett's Chi-Square Test
  - Scheffe Test

If the F-test assumptions are violated, then the F- distribution may not be the appropriate sampling distribution. A characteristic of ANOVA is that it is robust to violations of assumptions of the F-test as long as sample size, n, is equal. For example, Norton (1952) demonstrated robustness of the F-distribution to non-normality and heterogeneity of variance when equal sample size was used (referenced by Lindquist, 1956, pp. 78-86). Consequently, one should always strive to attain equal sample size across cells. If, however, one finds that most of the F ratios in an ANOVA are less than 1, this could be an indication of a marked violation of assumptions.

Transforming data to meet assumptions such as normality can be done. For example, latency data in human factors research is often positively skewed, and a log transformation can be used to normalize the data. The disadvantage of transformations is that the subsequent analysis is only valid for the transformed data and must be interpreted as such.

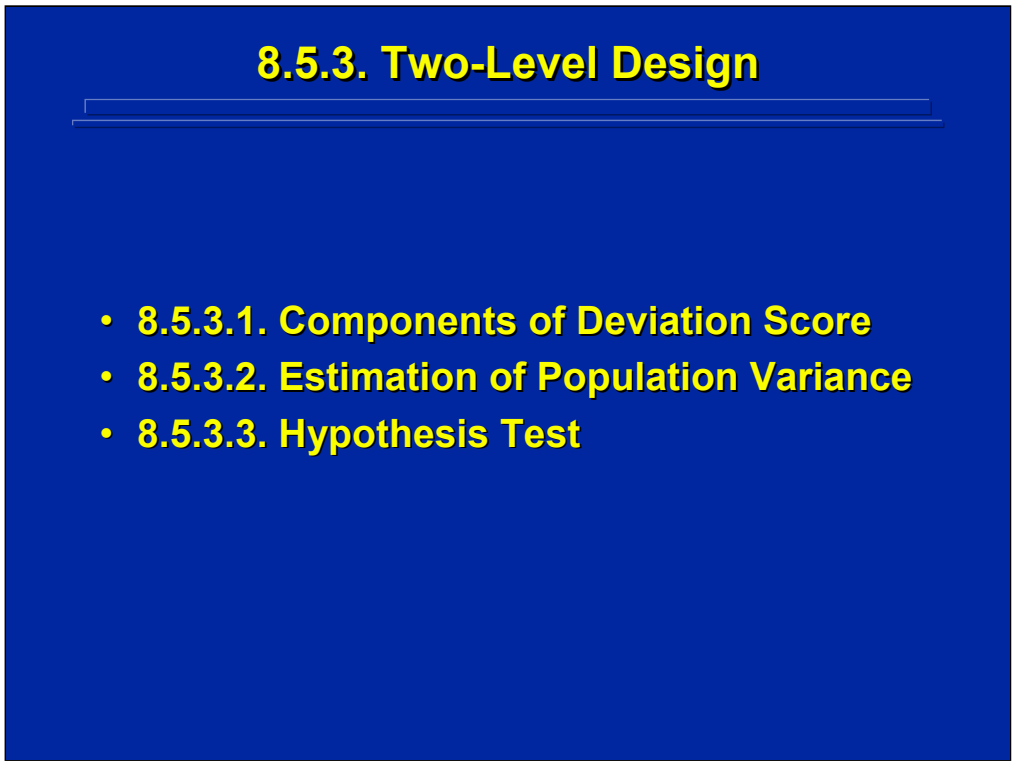
Violation of the homogeneity of variance assumption is critical when sample size is not equal. Four different alternatives for testing homogeneity of variance are described by Winer, Brown, and Michels (1991) on pp. 100-110. The Hartley F-Max Test is straightforward and often used. The ratio of the largest cell variance divided by the smallest cell variance in the data set provides the maximum value of F and is used in the F-Max Test as the observed value of F.

## 8.5.2. Assumptions of the F-Test (Cont'd)

- **Hartley F-Max Test**
  - $H_O: \sigma^2_1 = \sigma^2_2 = ... = \sigma^2_n$
  - $H_i: \sigma^2_1 \neq \sigma^2_2 \neq ... \neq \sigma^2_n$
  - $\alpha$: .20
  - D.R.: I reject $H_O$ if $F_{max} > F_{tab}$
    - $F_{max} = s^2_{Largest} / s^2_{Smallest}$
    - $F_{tab}$ = Table D.7 (Winer et al., 1991) where $n = n_{Largest}$
- **Heterogeneity of Variance**
  - Box Approximation to F (Winer, et al., 1991)

This slide shows the standard format for the Hartley F-Max Test. The null hypothesis states that the population variances are equal (homogeneity). The alternative hypothesis is that the population variances are not equal (heterogeneity). The decision rule is to reject the null hypothesis if the F-max statistic is greater than the F tabled value shown in Table D.7 of Winer, Brown, and Michels (1991). Usually, the F-Max Test is conducted at higher $\alpha$ error (i.e. $\alpha$ = .20) to guard against Type II error in accepting the null hypothesis (i.e., homogeneity of variance).

If the F-Max Test is significant, the homogeneity of variance assumption is violated and there is heterogeneity of variance. In this case the F distribution is not appropriate, and the Box approximation to F can be used (Winer, Brown, and Michels, 1991).




This subsection describes a simple F-test between two levels of one factor to demonstrate that ANOVA is really a test of differences between sample means even though the F-statistic is a ratio of sample variances. Deviation scores (i.e., the difference between an observation and its mean) that partition the variability of individual scores about the grand mean and two ways of estimating population variance are needed to demonstrate the overall logic of ANOVA hypothesis testing.

## 8.5.3. Two-Level Design (Cont'd)

- **One-Factor, Between-Subjects Data Set**

$$Y_{ijk} = \mu + \alpha_i + \gamma_{j(i)} + \varepsilon_{k(ij)}$$

| Night Vision Display A | Night Vision Display B |
|---|---|
| $Y_{11} = 59$ | $Y_{21} = 54$ |
| $Y_{12} = 65$ | $Y_{22} = 72$ |
| $Y_{13} = 52$ | $Y_{23} = 69$ |
| $Y_{14} = 45$ | $Y_{24} = 59$ |
| $Y_{15} = 63$ | $Y_{25} = 67$ |
| $Y_{16} = 42$ | $Y_{26} = 61$ |
| $Y_{17} = 53$ | $Y_{27} = 51$ |
| $Y_{18} = 47$ | $Y_{28} = 63$ |
| $Y_{1.} = 426$ | $Y_{2.} = 496$ |
| $\bar{Y}_{1.} = 53.25$ | $\bar{Y}_{2.} = 62.00$ |

Grand Total $Y.. = 922$

Grand Mean $\bar{Y}.. = 57.625$

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the basic layout of a two-group, between-subjects experiment that compared performance using night vision displays A and B by 8 different squads of soldiers. This experiment was described in the between-subjects t-test reviewed in Topic 3. As shown in the statistical model on the slide, this experiment can also be considered to be a simple one-factor, between-subjects ANOVA design in which the factor, Night Vision Display, has two levels, Display A and Display B.

The data set shows each of the 16 performance scores, $Y_{ij}$, where the subscript "i" refers to night vision display level, and subscript "j" refers to the 8 different squads of soldiers using each night vision display. Summing across levels of a factor is denoted by dotting the level designation of that factor in an observation. To compute a specific mean, the dotted sum is divided by the number of scores summed. For example, the total score for each night vision display type is denoted by dotting the "j" subscript (i.e., $Y_{1.}$ and $Y_{2.}$). These two treatment means are determined by $Y_{1.}/n$ and $Y_{2.}/n$, respectively. Likewise, the grand total of all 16 scores is shown by dotting both the "ij" subscripts (i.e., Y..), and the grand mean of the 16 scores is determined by Y../an.

## 8.5.3.1. Components of Deviation Score

- **Deviation Scores**
  - **Total = Within-Group + Between-Group**

$$(Y_{ij} - \bar{Y}_{..}) = (Y_{ij} - \bar{Y}_{i.}) + (\bar{Y}_{i.} - \bar{Y}_{..})$$

  - **(54 - 57.625) = (54 - 62) + (62 - 57.625)**
  - **(-3.625) = (-8) + (4.375)**
- **Sum of Squares**

$$SS_{Total} = SS_{Within\text{-}Group} + SS_{Between\text{-}Group}$$

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{i.})^2 + n\sum_{i=1}^{a}(\bar{Y}_{i.}-\bar{Y}_{..})^2$$

(Click in this red rectangle to see SAS calculations for this example.)

The total deviation of any score from the grand mean is the total of two additive parts made up of the within-group deviation and the between-group deviations, respectively. Within-group deviation is the difference between the observed value or score and its group mean. Between-group deviation is the difference between the group mean and the grand mean. The first observed value in the Night Vision Display B group, 54, is used on this slide as an example to demonstrate these deviation score relationships.

Deviation score relationships for an individual score also hold for sum of squared deviations around the mean (Myers 1979, pp. 76-83). Consequently, the $SS_{Total}$ equals $SS_{Within\text{-}Group}$ plus $SS_{Between\text{-}Groups}$. The formulae for calculating each of these sum of squares is shown on this slide in standard summation notation. These SS relationships are used to calculate the F-statistic in ANOVA.

## 8.5.3.2. Estimation of Population Variance

- **1. Pooled Estimate of Population Variance**
  - Mean Square

$$MS_{S/A} = \sum_{i=1}^{a} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i.})^2 / a(n-1)$$

  - Treatment Means Do Not Affect Estimate
- **2. Sampling Distribution of Means**
  - Mean Square

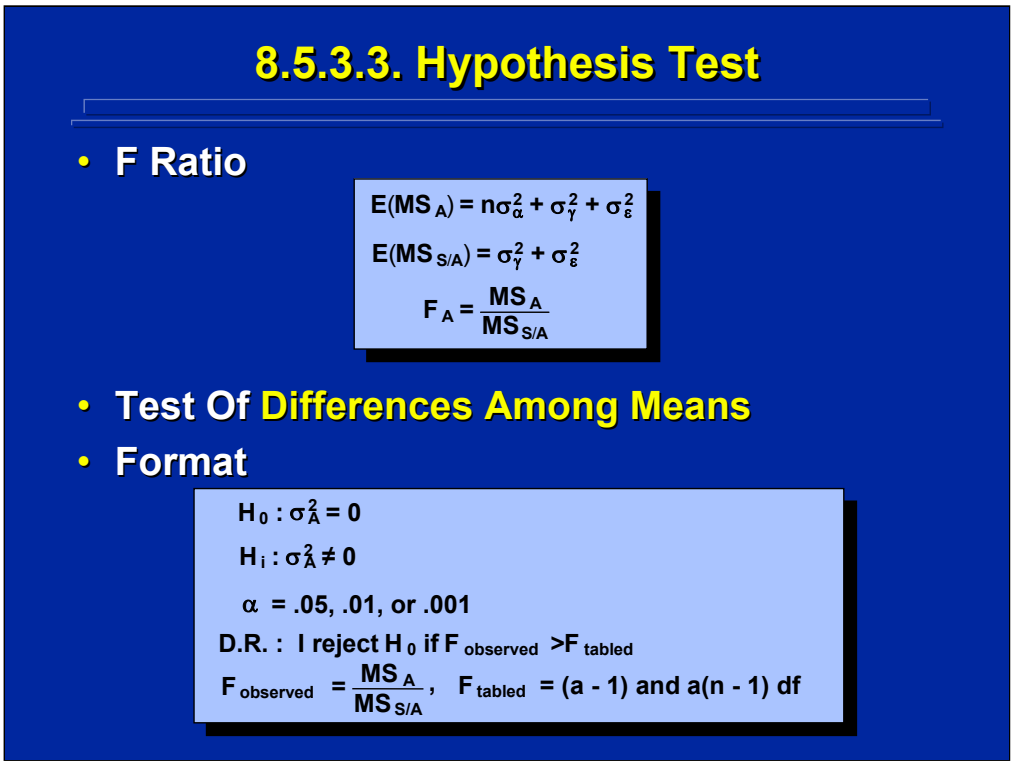
$$MS_A = n \sum_{i=1}^{a} (\bar{Y}_{i.} - \bar{Y}_{..})^2 / (a-1)$$

  - Treatment Means Do Affect Estimate

(Click in this red rectangle to see SAS calculations for this example.)

Mean square is another term for variance and is equal to a sum of squares (SS) divided by its degrees of freedom (df). Two independent estimates of population variance are calculated in an F-statistic using SS and df. The first estimate of population variance is calculated from the within-group SS in which the sum of squares is pooled across groups and is referred to as $MS_{S/A}$ since this is a between-subjects design. Since the deviations of individual scores from their group mean are calculated separately and then pooled across groups, the treatment means do not affect this pooled estimate of population variance.

The second estimate shown on this slide is referred to as $MS_A$ and is calculated from between-group SS. In this case, the population variance is estimated from the sampling distribution of means. Since group mean deviations are calculated from the grand mean, the treatment means do affect this estimate of population variance.




(Click in this red rectangle to see SAS calculations for this example.)

The top portion of this slide shows the theoretical components that can affect the estimates of population variance calculated by $MS_A$ and $MS_{S/A}$. Since $MS_A$ is affected by treatment means, it has a theoretical treatment component ($\alpha$) in addition to the subjects ($\gamma$) and random error ($\varepsilon$) components. On the other hand, $MS_{S/A}$ is affected theoretically only by the subject ($\gamma$) and random error ($\varepsilon$) components. Consequently, to obtain a theoretical F = 1 under the null hypothesis, one would use the ratio of $MS_A$ divided by $MS_{S/A}$. If the resulting F-test is significant, then there is a significant difference between treatment means.

The standard format for performing any F-test is shown on the bottom portion of this slide. The experimenter calculates the observed value of F and compares it to the tabled values from the F-statistic sampling distribution. The same logic used to determine the numerator and denominator of the observed F-ratio in a one-factor design is followed in complex ANOVA designs that include several factors whether they are between-subjects, within-subjects, or mixed-factors designs.

## 8.5.3.3. Hypothesis Test (Cont'd)

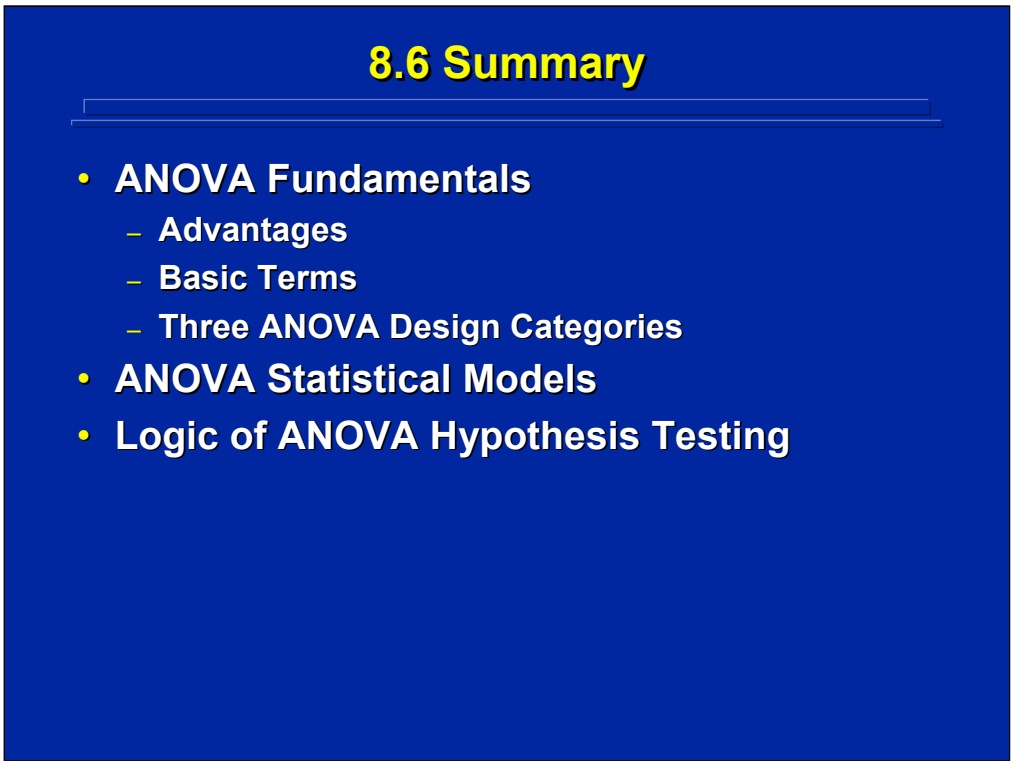
- **Night Vision Display Example**
  - $MS_A = 306.25$
  - $MS_{S/A} = 62.25$
  - $F_A = 306.25/62.25 = 4.92$
- **Standard Format**
  - $H_0: \sigma^2_A = 0$
  - $H_1: \sigma^2_A \neq 0$
  - $\alpha = .05$
  - D.R.: I reject $H_0$ if $F_{Observed} > F_{Tabled}$
  - where $F_{Observed} = 4.92$ and $F_{(1,14)} = 4.60$
- **t-Test Result**
  - $t_{Observed} = 2.22$ $(t^2_{Observed} = 4.92)$

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the results of using the formulae presented for $MS_A$ and $MS_{S/A}$ to calculate the actual values for the two mean squares as well as the resulting F-ratio for the data set of the night vision example. Slater and Williges (2006) present the SAS program for calculating the complete two-level, one-way ANOVA.

The standard format for testing mean performance differences between the two types of night vision displays is shown in the middle portion of this slide. Note that the tabled value of the F-ratio shows the degrees of freedom of the numerator and denominator of the F-ratio in parenthesis. Since the observed value of the F-ratio is greater than the tabled value of the F-ratio, one concludes that there is a significant difference between the two night vision displays at the 0.05% level of significance.

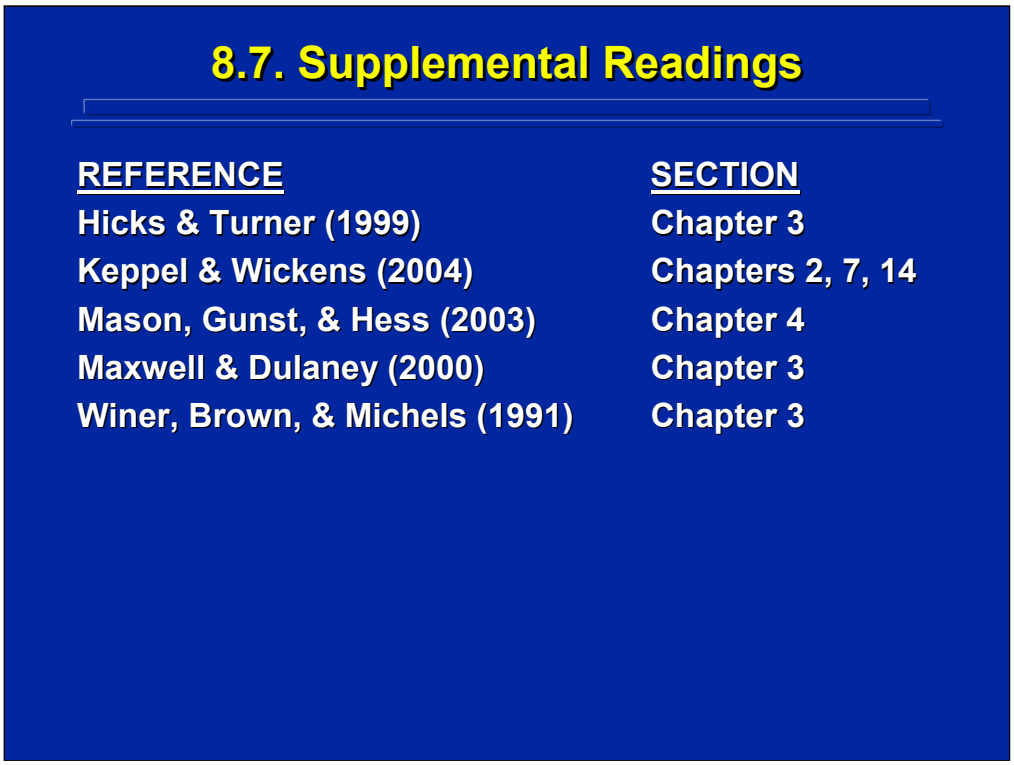
The bottom portion of this slide shows the results of the between-subjects t-test conducted on the night vision display as discussed in Topic 3. Recall that $(t_\nu)^2 = F_{1,\nu}$ as demonstrated in the results of this example that are presented on the slide. So, a two-level, one-factor ANOVA is equivalent to a t-test of the difference between the means of two groups.




By way of summary, this introductory topic on ANOVA presented the three main concepts shown on this slide. For the human factors and ergonomics researcher, ANOVA designs have the advantage of including many factors simultaneously in one experiment to evaluate the main effects and interactions among them. Basic terms were presented that provide the essential vocabulary for describing any ANOVA design. In human factors research, ANOVA designs can be categorized as either between-subjects, within-subjects, or mixed-factors designs depending the crossing and nesting of subjects with factors of interest.

The basic fabric of every ANOVA design is defined by the statistical model. The ANOVA statistical model states that every observation in a human factors experiment is conceptually a linear combination of the population mean, treatments, subjects, interactions, and random error effects. Steps for specifying statistical models were provided. The researcher must be able to specify the statistical model in order to define the experimental design and specify the effects that can be evaluated in the experiment.

Finally, the overall logic for conducting ANOVA hypothesis testing on difference between means was presented. An example of using this logic in a simple two-level, one-factor ANOVA design showed the standard format for any ANOVA hypothesis testing.



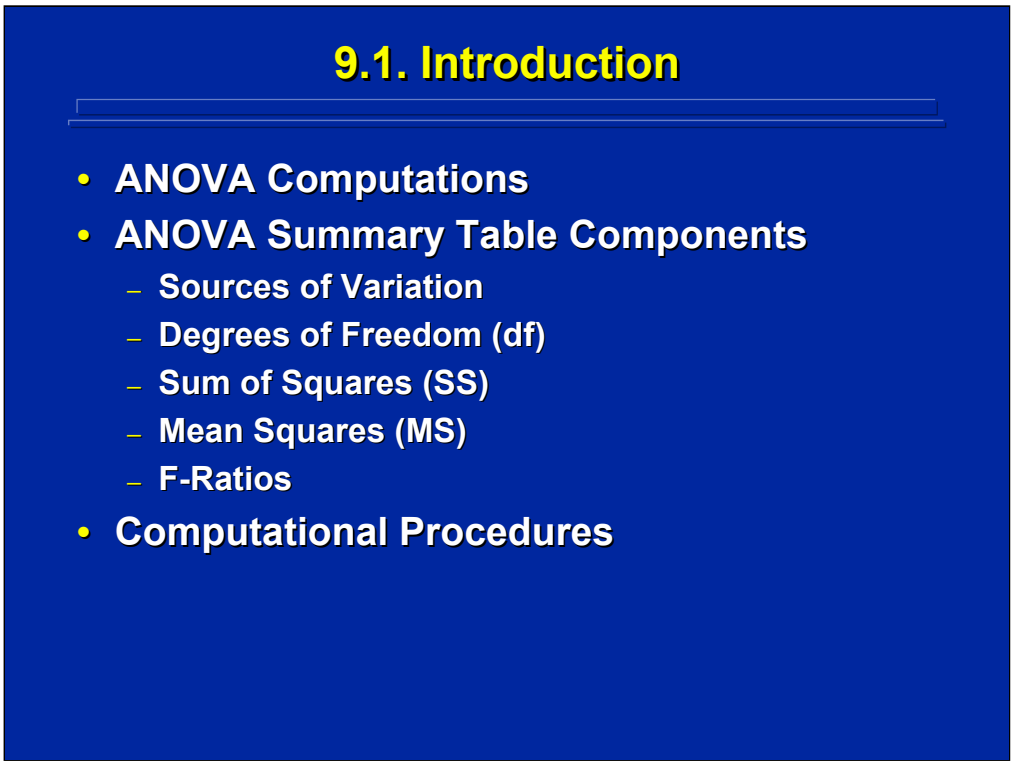

All these texts provide general introductions to ANOVA designs and a detailed description of the logic involved in ANOVA hypothesis testing. In addition, Keppel and Wickens (2004) provide an overview of the use of linear models in ANOVA in Chapter 7 as opposed to the variance component statistical models discussed in this chapter. Winer, Brown, and Michels (1991) provide a detailed description of testing for homogeneity of variance in ANOVA on pages 100-110.

# Topic 9. ANOVA Summary Table Components

9.1. Introduction
9.2. Sources of Variation
9.3. Degrees of Freedom (df)
9.4. Sum of Squares (SS)
9.5. Mean Squares (MS)
9.6. F-Ratios
9.7. Complete ANOVA Summary Table
9.8. ANOVA Design Construction
9.9. Summary
9.10. Supplemental Readings

This topic provides an overview of the computational aspects of any between-subjects, within-subjects, and mixed-factors ANOVA design used in human factors research. The subsections are organized around the five major components of the ANOVA Summary Table used for listing the results of an ANOVA. Rather than derive formulae for calculating each component, computational procedures and algorithms are provided.

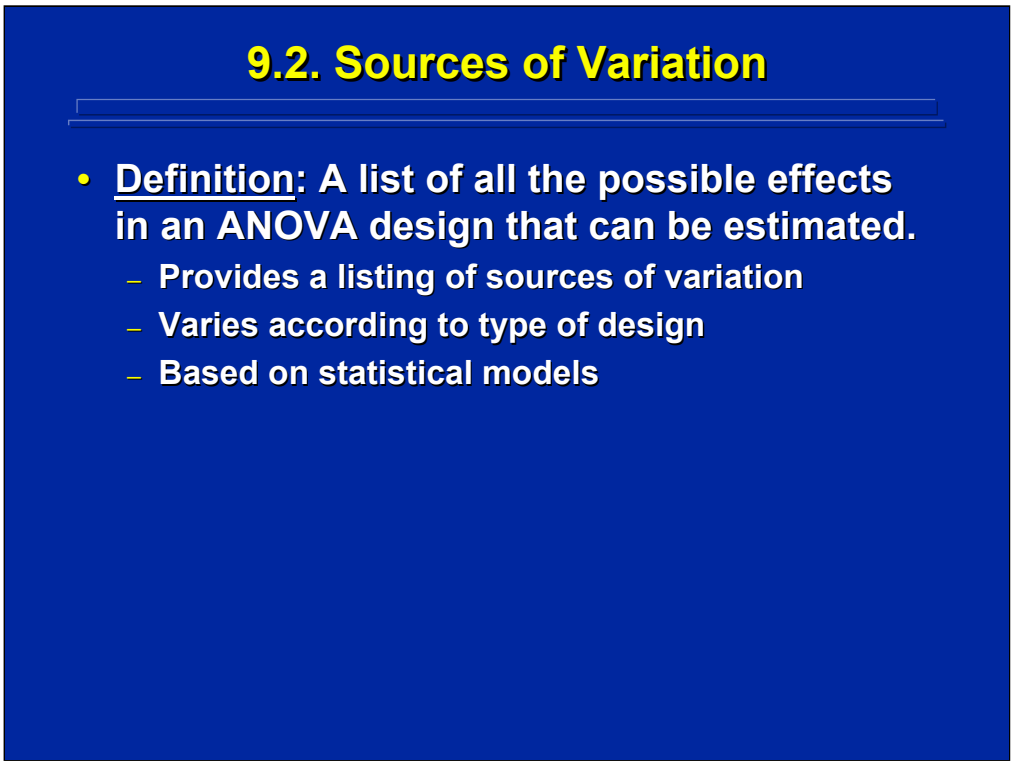
First, each component of the summary table is discussed separately. Then conventions for stating the complete ANOVA Summary Table are presented for each of the three major categories of ANOVA experimental designs used in human factors research. References to supplemental readings on ANOVA Summary Table details are provided for the major experimental design texts appropriate for human factors research.



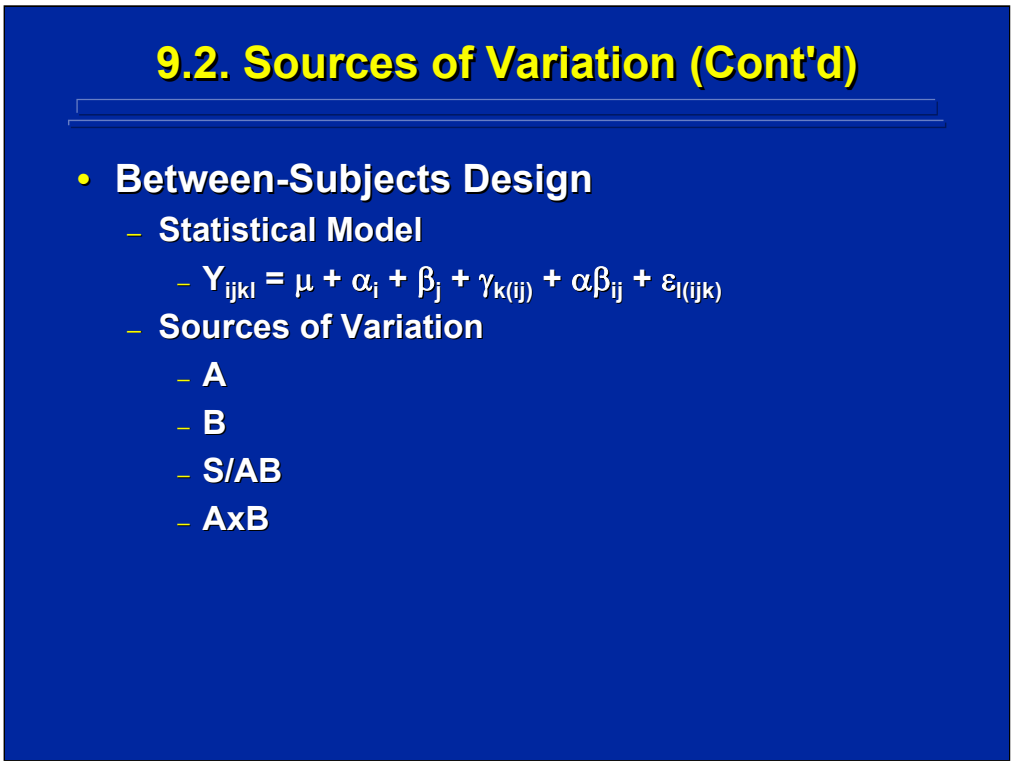

Due to the computational complexity of ANOVA, most human factors researchers use statistical analysis packages for conducting the ANOVA on complex designs. But, the researcher needs to understand the ANOVA computations in order to check the accuracy of the statistical analysis program output. This topic provides an overview of the various analysis components of ANOVA for any factorial design used in human factors and ergonomics research.

The five major components of an ANOVA Summary Table are presented in the center portion of this slide. Each component is discussed separately, analogous to describing the pieces of a puzzle. The complete ANOVA Summary Table is presented for between-subjects, within-subjects and mixed-factors designs to summarize the relationship of these components in statistical hypothesis testing. Actual computations of the ANOVA are presented in Topic 9 following this general discussion of computational procedures.

Topic 8 provided the general logic of an ANOVA hypothesis test using deviation scores and definitional formulae. Rather than deriving formulae for each possible design separately, this topic presents general rules, procedures, and algorithms that provide the same results as derivation, but are easier to use and generalize across ANOVA designs.

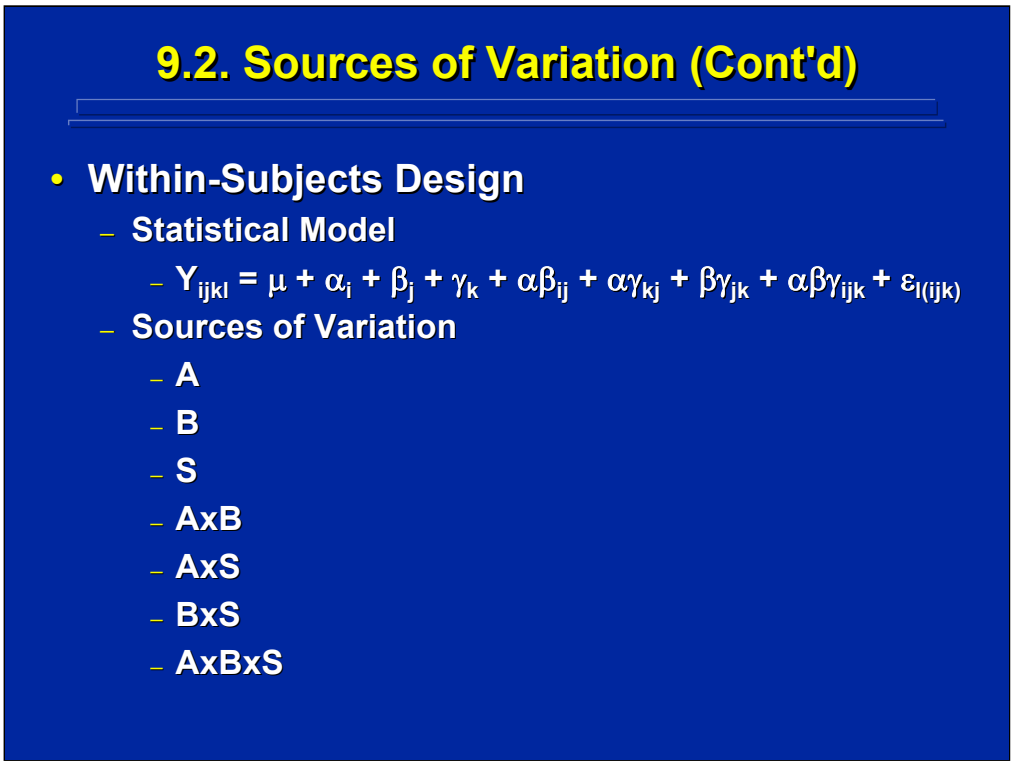



The first component in an ANOVA Summary Table is a listing of sources of variation. The Source listing provides all the treatment components of the experimental design that can be estimated from the data set and are involved in subsequent ANOVA calculations. The possible sources vary according to the particular ANOVA experimental design and are based directly on the statistical model for that design. Sources for between-subjects, within-subjects, and mixed-factors designs are presented separately using a two-factor ANOVA design as an example.

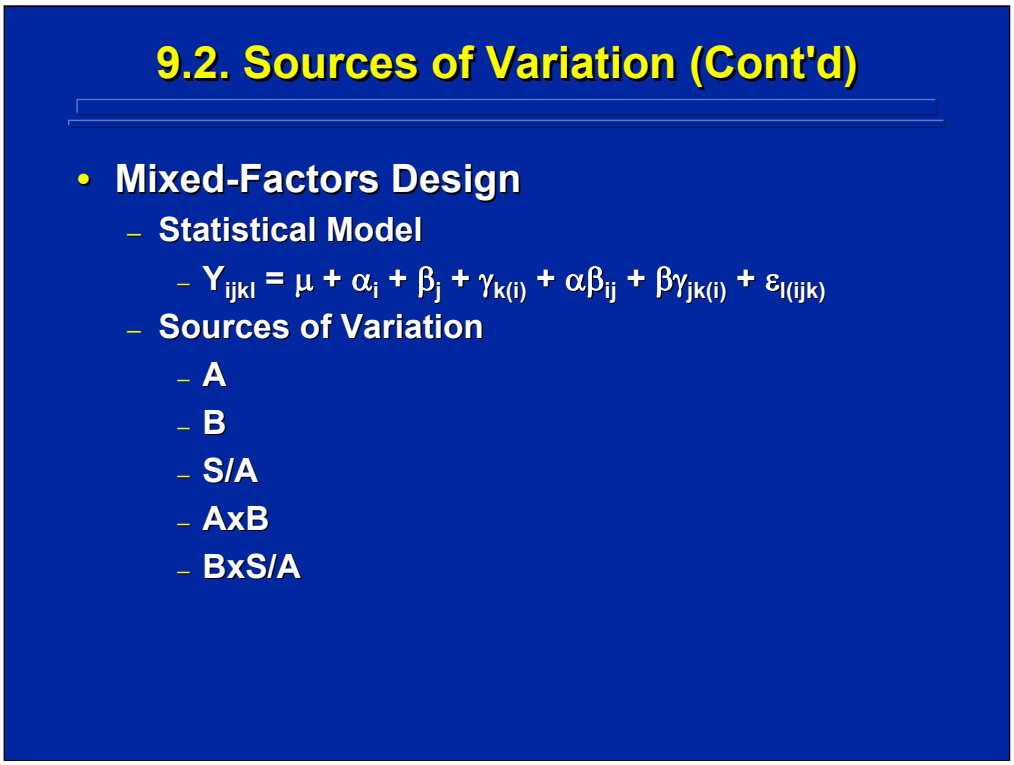



The statistical model for a two-factor, between subjects ANOVA design is shown in the top portion of this slide. All the treatment components in this design are listed between the population mean, $\mu$, and random error, $\varepsilon$, in the statistical model. Specifically in this example, there are 4 treatment components.

Based on the nesting of effects listed in the statistical model, one can determine the appropriate listing of the sources of variation calculated in this ANOVA design. The resulting sources of variation for this particular design as stated in standard notation are A, B, S/AB, and AxB. Note that subjects are nested in both factors A and B since this is a between-subjects experimental design.

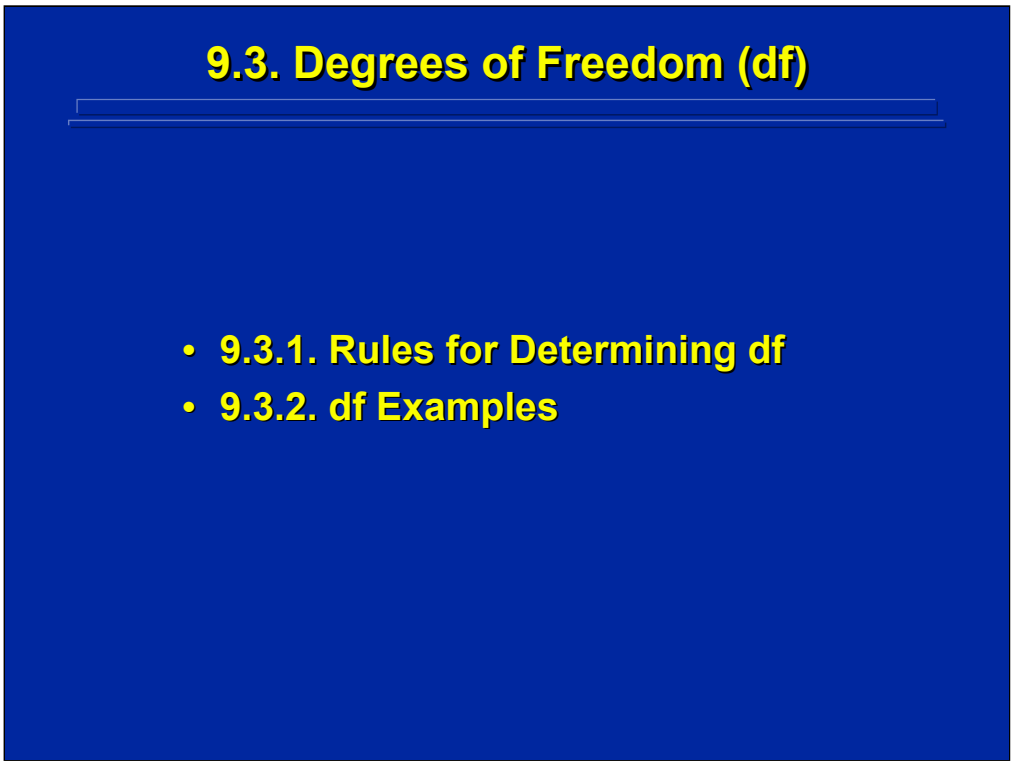



The statistical model for the within-subjects version of the two-factor ANOVA design is shown on the top portion of this slide. Since subjects are crossed with all factors of interest, everything can interact resulting in seven sources of treatment variation. The sources of variation for this design are A, B, S, AxB, AxS, BxS, AxBxS. There are three main effects, three two-way interactions, and one three-way interaction that are calculated in the ANOVA for this design.

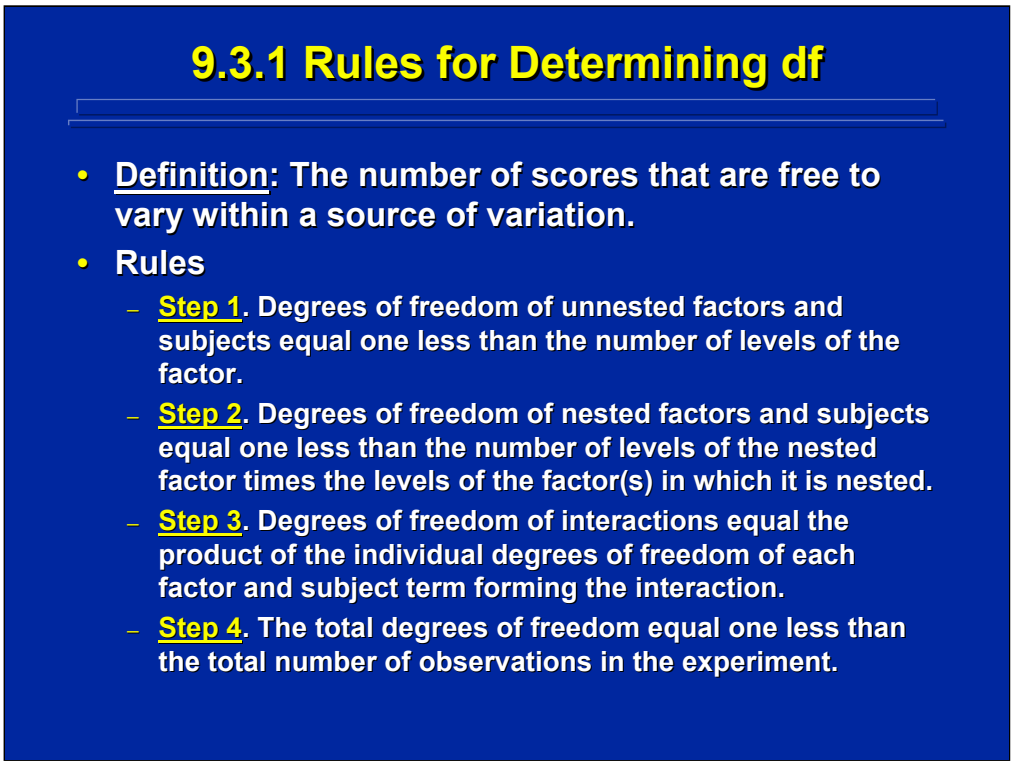



A mixed-factors version of the two-factor design is illustrated on this slide. Based on the nesting shown for $\gamma$ in the statistical model Subjects are nested in Factor A and crossed with Factor B. This ANOVA design results in 5 sources of variation that are evaluated in the subsequent ANOVA. The sources of variation for this design are A, B, S/A, AxB, and BxS/A.

These three examples of a two-factor ANOVA design demonstrate that the possible sources of variation in the ANOVA Summary Table vary depending upon whether the design is a between-subjects, within-subjects, or mixed-factors design. In all cases, however, the sources of variation can be determined directly from the statistical model for the design. All three design alternatives include the three sources of overall interest to the experiment (e.g., Factor A, Factor B, and the AxB interaction) and differ only in effects due to subjects. Obviously, this procedure for determining the Source listing generalizes to any number of factors of interest included in the ANOVA experimental design.




The second component of an ANOVA Summary Table lists the degrees of freedom associated with each source of variation in the experiment. This subsection lists some simple rules for determining df and provides examples of applying these rules to between-subjects, within-subjects, and mixed-factors, two-way ANOVA designs.




Degrees of freedom are the number of scores that are free to vary within the various sources of the design. In general, every time you calculate a statistic you lose one degree of freedom. For example, if a dataset is composed of 24 numbers, the mean can be equal to any value if one number is fixed and the other 23 numbers are free to vary. Consequently, the df for the grand mean is one less than the total number of observations or 23 as stated in Step 4 on the slide.

The lower portion of this slide provides four simple steps for determining the df of any source of variation in ANOVA assuming sample size, n, is equal in each cell of the design. These four steps are patterned after Keppel and Wickens (2004, p.215). The first step deals with unnested sources, and the second step deals with nested sources. The third step pertains to interactions. Finally, the fourth step specifies the total df in an ANOVA design.

Since df are additive in ANOVA, the sum of the df for all sources in an experiment should equal the total df in a design. This provides a simple check that all sources of variation are included in the ANOVA Summary Table.

## 9.3.2. df Examples

- 3x2 Between-Subjects Design, n=4

| Sources | Degrees of Freedom |
|---|---|
| A | (a-1) = 2 |
| B | (b-1) = 1 |
| S/AB | ab(n-1) = (3)(2)(3) = 18 |
| AxB | (a-1)(b-1) = (2)(1) = 2 |
| Total | abn-1 = (3)(2)(4) = 23 |

Consider a two-factor ANOVA design in which Factor A has 3 levels (i.e., a=3), Factor B has 2 levels (i.e. b=2), and 4 subjects are observed in each cell of the factorial design (i.e., n=4).

The df for the between-subjects alternative for this 3x2 factorial design is shown on this slide. The df for Factors A and B are determined by Step 1 in the rules for determining df. Since subjects are completely nested in this design, the df for S/AB are determined by Step 2. The AxB interaction is determined by Step 3. Finally, the total degrees of freedom in this two-factor design is 23 as determined by Step 4 of the rules for determining df.

Note that the df of all the Sources sum to the total df in the design. Always calculate the df for each source directly and do not determine any of them by subtraction. Comparison of the sum of all sources to the calculated total df provides a simple check that all the sources of variation are included and no mistake was made in determining the df for one or more of these sources.

## 9.3.2. df Examples (Cont'd)

- 3x2 Within-Subjects Design, n=4

| Sources | Degrees of Freedom |
|---|---|
| A | (a-1) = 2 |
| B | (b-1) = 1 |
| S | (n-1) = 3 |
| AxB | (a-1)(b-1) = (2)(1) = 2 |
| AxS | (a-1)(n-1) = (2)(3) = 6 |
| BxS | (b-1)(n-1) = (1)(3) = 3 |
| AxBxS | (a-1)(b-1)(n-1) = (2)(1)(3) = 6 |
| Total | abn-1 = (3)(2)(4) = 23 |

The df for the within-subjects design alternative for the 3x2 factorial design is shown on this slide. Since there is no nesting in this design, Rule 2 for determining df does not apply. Note that the total df still equals 23, and the df for all sources of variation in the experiment sums to 23.

## 9.3.2. df Examples (Cont'd)

- 3x2 Mixed-Factors Design, n=4

| Sources | Degrees of Freedom |
|---|---|
| A | (a-1) = 2 |
| B | (b-1) = 1 |
| S/A | a(n-1) = (3)(3) = 9 |
| AxB | (a-1)(b-1) = (2)(1) = 2 |
| BxS/A | a(b-1)(n-1) = (3)(1)(3) = 9 |
| Total | abn-1 = (3)(2)(4) = 23 |

The df for a 3x2 mixed-factors design is shown on this slide. The four rules for determining df still apply regardless of the partial nesting of subjects in the experiment. Note that the df for S/A are determined by Rule 2. Subjects are only nested in A and the df are affected by the levels of A, but the levels of Factor B are not involved in determining the interaction. Also note that mixed factor designs have interactions involving nesting (e.g. BxS/A). Rule 3 for determining the df of interactions still apply. One simply uses the degrees of freedom for each factor component of the interaction (e.g. 1x9=9)

Consider all three of these 3x2 design examples. The sources change due to the crossed versus nesting relations between Subjects and both Factors A and B. The total df in all three experiments is 23, and the df for the three effects of interest to the experiment remain the same (e.g. A=2, B=1, and AxB=2).

## 9.4. Sum of Squares (SS)

- **Definition: The sum of squared deviations around the mean is the actual observed deviation of each source.**
- **Definition Form**

$$SS = \Sigma (Y_i - \bar{Y})^2$$

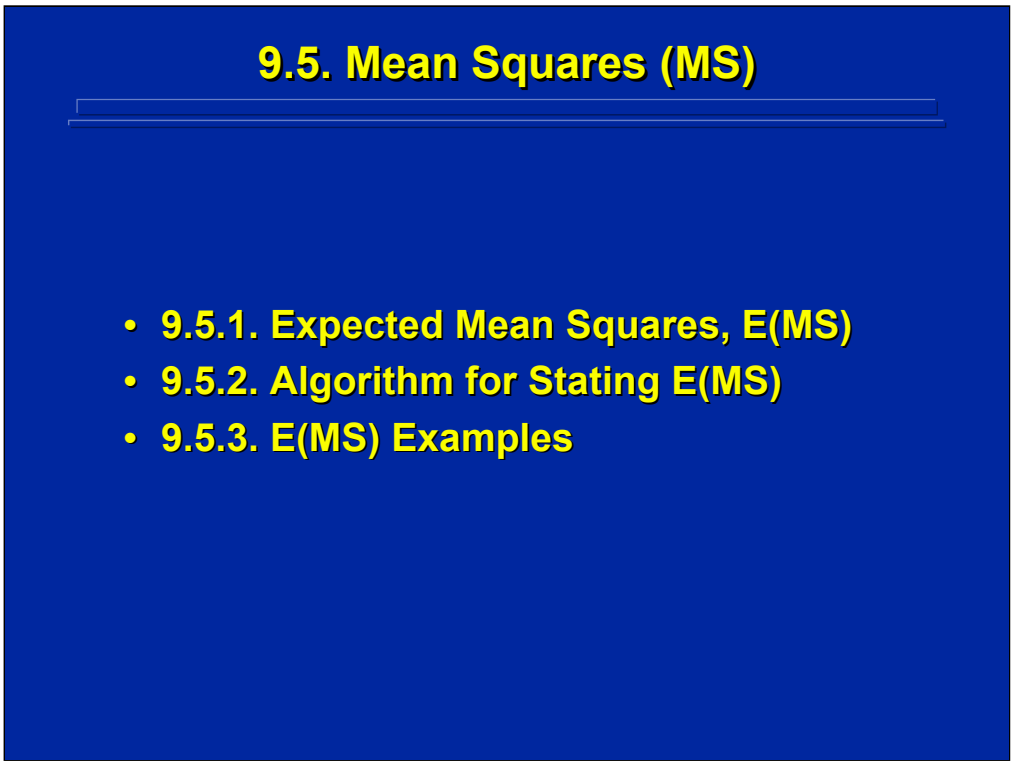
- **Computational Form**

$$SS = \Sigma Y_i^2 - [(\Sigma Y_i)^2 / n]$$

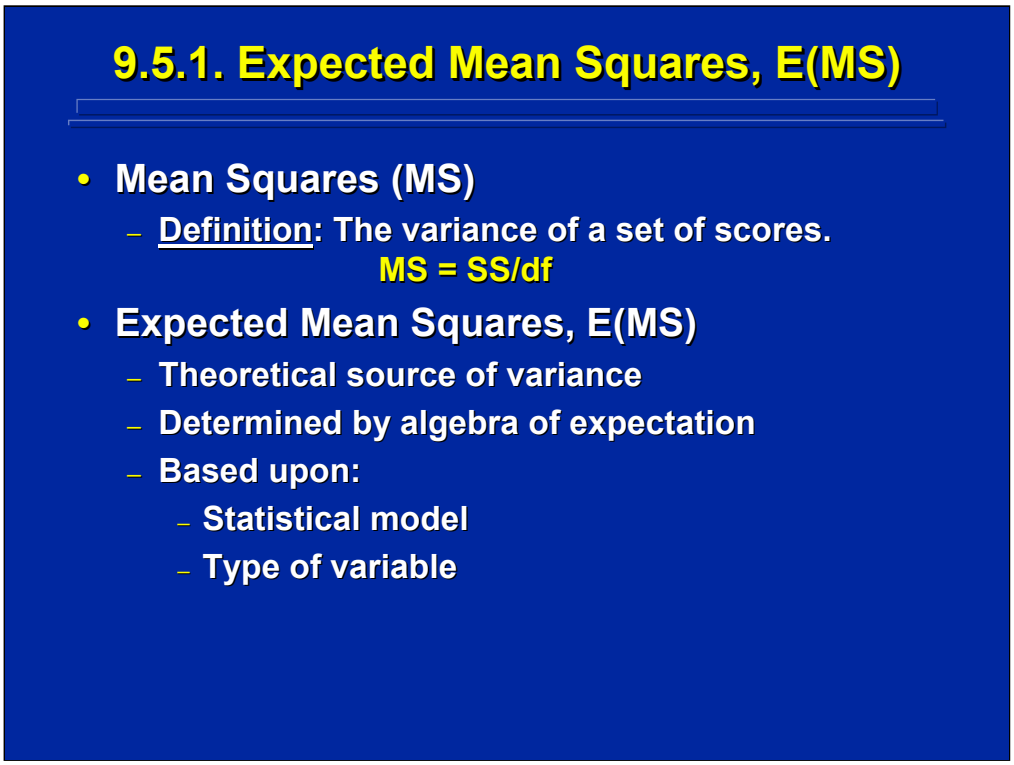
- **SS computational formulae are specific for each source of variability.**

Sum of Squares (SS) is the third component of an ANOVA Summary Table. The SS is really the sum of the squared deviations around the mean which is based on the actual observed deviation of each score from its appropriate cell mean.

The computational form for the SS uses raw scores to avoid calculations of an intermediate mean as shown on this slide. Sum of squares computational formulae are specific for each source of variability. Rather than state the definitional formula for the SS of each source of variation and then convert each definitional formula into its computational form, an algorithm is provided in Topic 9 that specifies the various computational formulae directly. Since calculations of SS are the major analytical component of ANOVA, details on this procedure will be discussed in Topic 9 when a complete ANOVA computational example is discussed. Consequently, formulae are not provided for each source of variation in this topic but are only designated as sums of squares with the appropriate subscripts (e.g., $SS_A$, $SS_B$, and $SS_{AXB}$) in this topic.

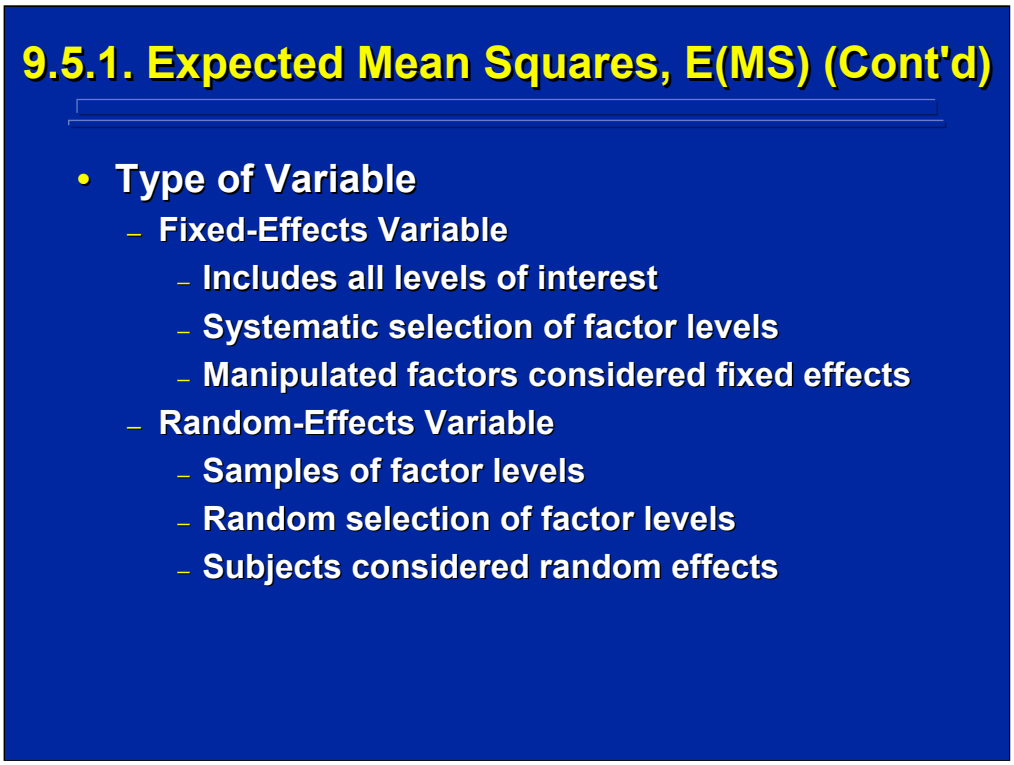



Mean squares (MS) provide the fourth component of the ANOVA Summary Table and represent variance calculations that are used in constructing F ratios. In order to determine the appropriate MS to use in the denominator of any F ratio, one must consider the theoretical sources of variance that are represented in a MS. These theoretical components are called expected mean squares, E(MS). This subsection, discusses factors that determine E(MS), describes an algorithm for casting E(MS), and provides examples of E(MS) in the three major categories of human factors ANOVA designs.




As shown in the top portion of this slide the MS or variance of a set of scores is nothing more than the SS of the scores divided by the df. In an ANOVA Summary Table, one can easily calculate the MS for any source of variation by dividing its SS by its df.

The theoretical sources of variance, or the E(MS), that comprise a MS are determined through the algebra of expectation. Two parameters, the statistical model and the type of variable manipulated in the experiment, determine the components of an E(MS). The statistical model of the design lists all the components of variation that are present in an observed score. The type of variable determines the effect of a factor on the observation.



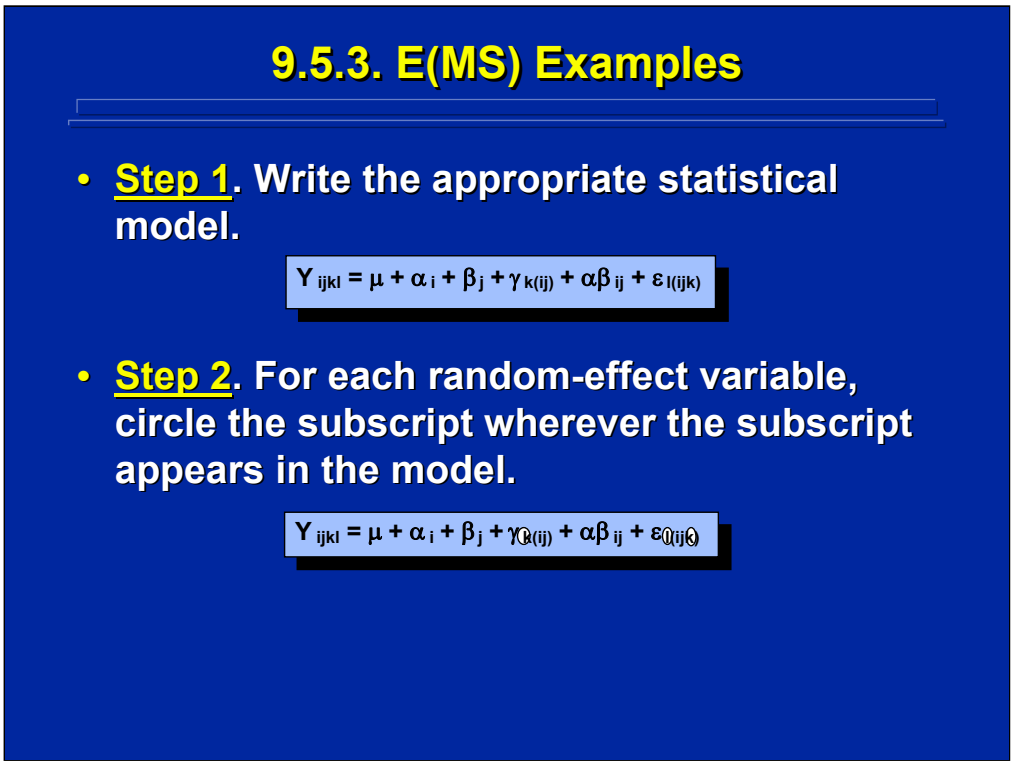

Any variable in an experiment can be classified as either a fixed-effects variable or a random-effects variable. A fixed-effects variable means that the experimenter has systematically selected all of the factor levels that exist for that factor. The random-effects variable means that the experimenter has made only a random selection of possible levels of a variable.

In human factors experiments, factors of interest are considered fixed-effects variables because the experimenter includes all factor levels of interest to the experiment. Generalizations of results of the experiment are, in turn, restricted to those levels of the factor. Subjects, on the other hand, is considered a random-effects variable because the experimenter makes a random selection of possible subjects to participate in the experiment. The results of the experiment generalize to the population of subjects from which the random sample was drawn. Random selection must occur to have a truly random-effects variable. Consequently, manipulated factors in an experiment are considered fixed-effects variables and subjects are considered a random-effects variable when constructing E(MS).

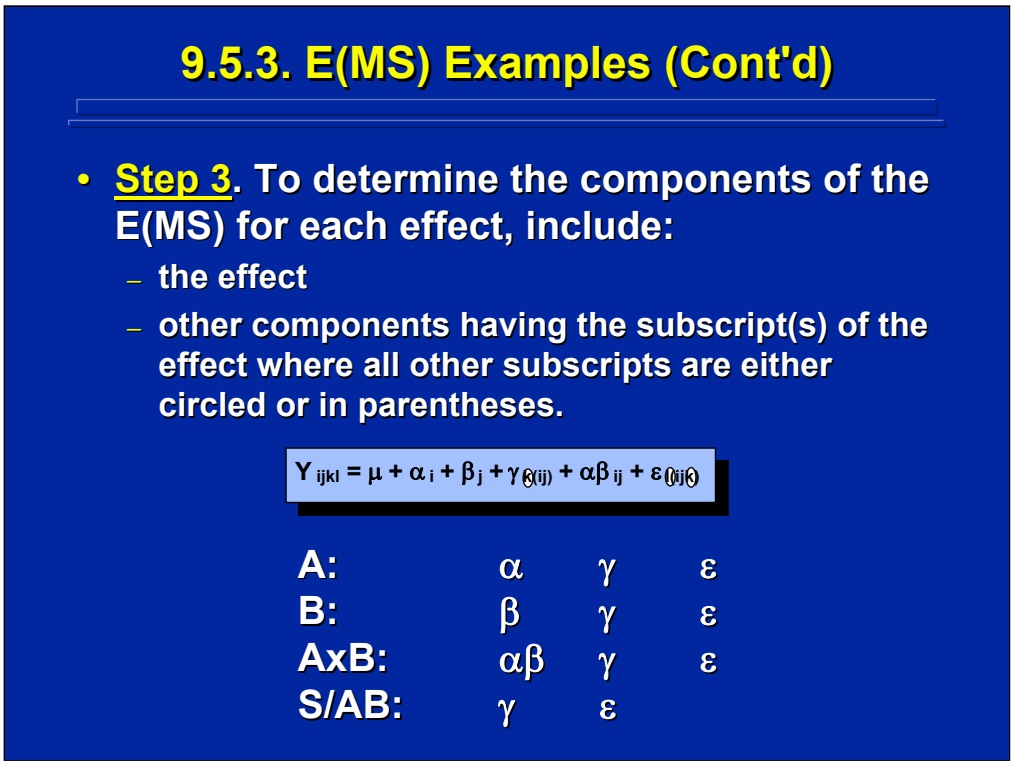
## 9.5.2. Algorithm for Stating E(MS)

- Step 1. Write the appropriate statistical model.
- Step 2. For each random-effect variable, circle the subscript wherever the subscript appears in the model.
- Step 3. To determine the components of the E(MS) for each effect, include:
  - the effect; and
  - other components having the subscript(s) of the effect where all other subscripts are either circled (random effect) or in parentheses (nested).
- Step 4. Begin to list the E(MS) for each effect as a linear combination of the $\sigma^2$ for each component. Note that the subscript for each $\sigma^2$ is the Greek symbol(s) of the component.
- Step 5. To complete the E(MS) listing, multiply each $\sigma^2$ in the resulting linear combination by the number of levels of the factor(s) not involved in defining the component term.

The weighted components of an E(MS) are derived through the algebra of expectations, and this derivation can become tedious. See Winer, Brown, and Michels (1991, pp. 89-100) for a discussion of these mathematical procedures. Alternatively, an algorithm can be used for stating the E(MS) in lieu of actual derivation through the algebra of expectation. In addition, Montgomery (2005, pp. 501-505) and Myers and Well (2003, pp. 392-394) and Winer, Brown, and Michels (1991, pp. 369-374) present alternative rules for generating E(MS) that provide essentially the same results. Subsequent slides demonstrate the use of this algorithm for a two-factor between-subjects design and show the resulting E(MS) for both a two-factor within-subjects and mixed-factors design.




A two-factor, between-subjects design is used as an example to demonstrate the 5-step algorithm for casting E(MS). In Step 1, the complete between-subjects statistical model is listed. In Step 2, the subscripts of the two random-effects factors, $\gamma$ and $\varepsilon$, are circled whenever they appear in the statistical model. Both $\alpha$ and $\beta$ are fixed-effects factors, and their subscripts are not circled.




As shown on this slide, Step 3 provides an initial listing of components that are included in the E(MS) of each source of variance listed in the ANOVA Summary Table. First, the component of the statistical model representing the source is listed. Next, other components that include the subscript of that source are included provided all other subscripts are either circled or in parentheses. For example, $\alpha\beta_{ij}$ is not a component of A since the subscripts "ij" are not circled or in parentheses. Notice that the source effect, the subjects effect, and the random error effect are the only contributors to the E(MS) for each source of variance in this between-subjects design.

## 9.5.3. E(MS) Examples (Cont'd)

- **Step 4. Begin to list the E(MS) for each effect as a linear combination of the $\sigma^2$ for each component. Note that the subscript for each $\sigma^2$ is the Greek symbol(s) of the component.**

$$E(MS_A) = \sigma^2_\alpha + \sigma^2_\gamma + \sigma^2_\varepsilon$$
$$E(MS_B) = \sigma^2_\beta + \sigma^2_\gamma + \sigma^2_\varepsilon$$
$$E(MS_{AxB}) = \sigma^2_{\alpha\beta} + \sigma^2_\gamma + \sigma^2_\varepsilon$$
$$E(MS_{S/AB}) = \sigma^2_\gamma + \sigma^2_\varepsilon$$

Step 4 of the algorithm for casting E(MS) merely takes the components from Step 3 and lists each of them as a subscript of a variance contributor, $\sigma^2$, in a linear combination.

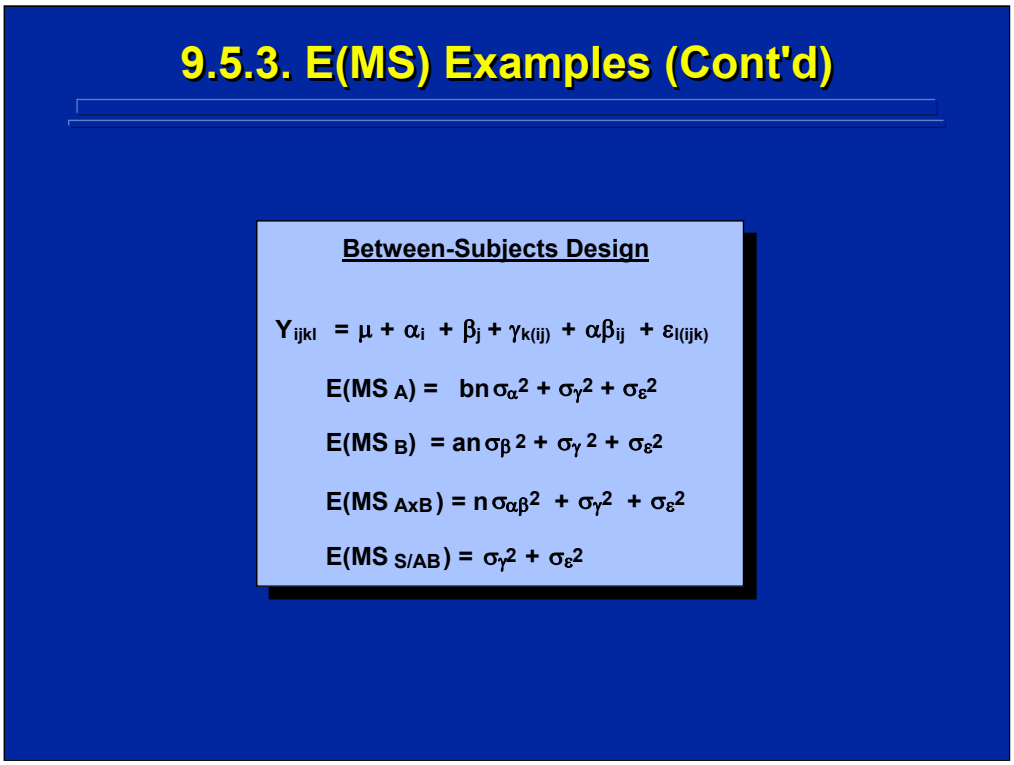
## 9.5.3. E(MS) Examples (Cont'd)

- **Step 5. To complete the E(MS) listing, multiply each $\sigma^2$ in the resulting linear combination by the number of levels of the factor(s) not involved in defining the component term**

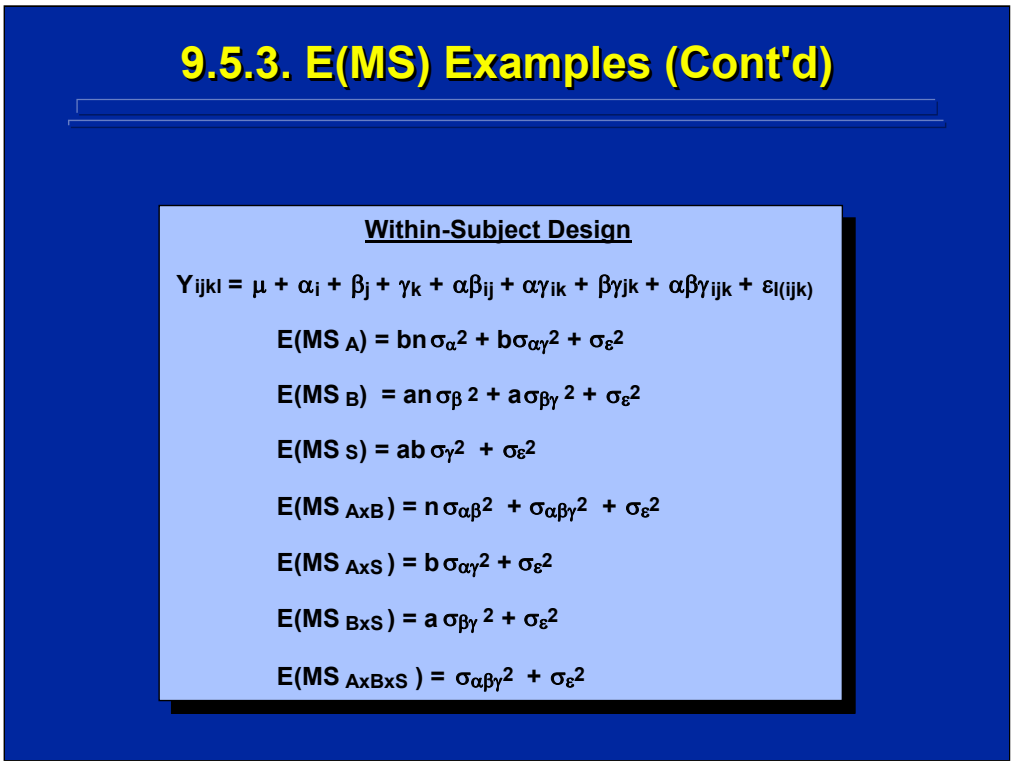
$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_{k(ij)} + \alpha\beta_{ij} + \varepsilon_{l(ijk)}$$

$$E(MS_A) = bn\sigma^2_\alpha + \sigma^2_\gamma + \sigma^2_\varepsilon$$
$$E(MS_B) = an\sigma^2_\beta + \sigma^2_\gamma + \sigma^2_\varepsilon$$
$$E(MS_{AxB}) = n\sigma^2_{\alpha\beta} + \sigma^2_\gamma + \sigma^2_\varepsilon$$
$$E(MS_{S/AB}) = \sigma^2_\gamma + \sigma^2_\varepsilon$$

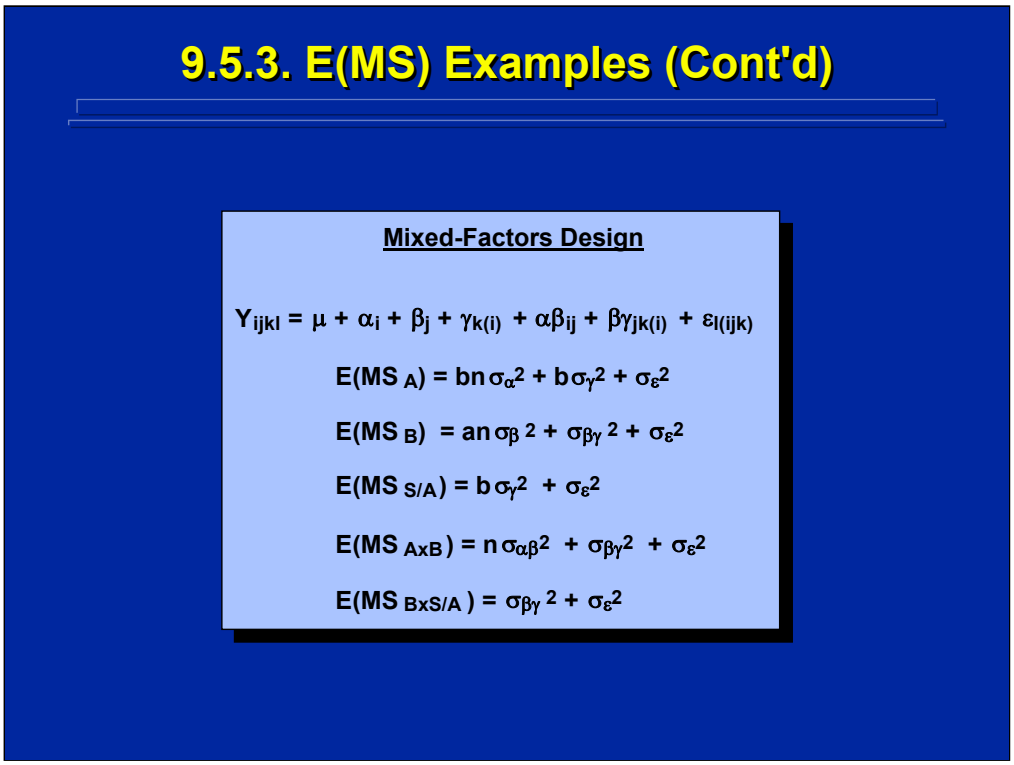
The final step in the algorithm is the determination of weightings for each contributor in the E(MS). These weightings are simply the number of levels of all the factors NOT included as a subscript of that effect in the statistical model. This always results in a weighting of 1 for $\varepsilon$ since it always includes every subscript. The final representation of E(MS) for each source of variance in this between-subjects design is shown in the bottom portion of this slide.



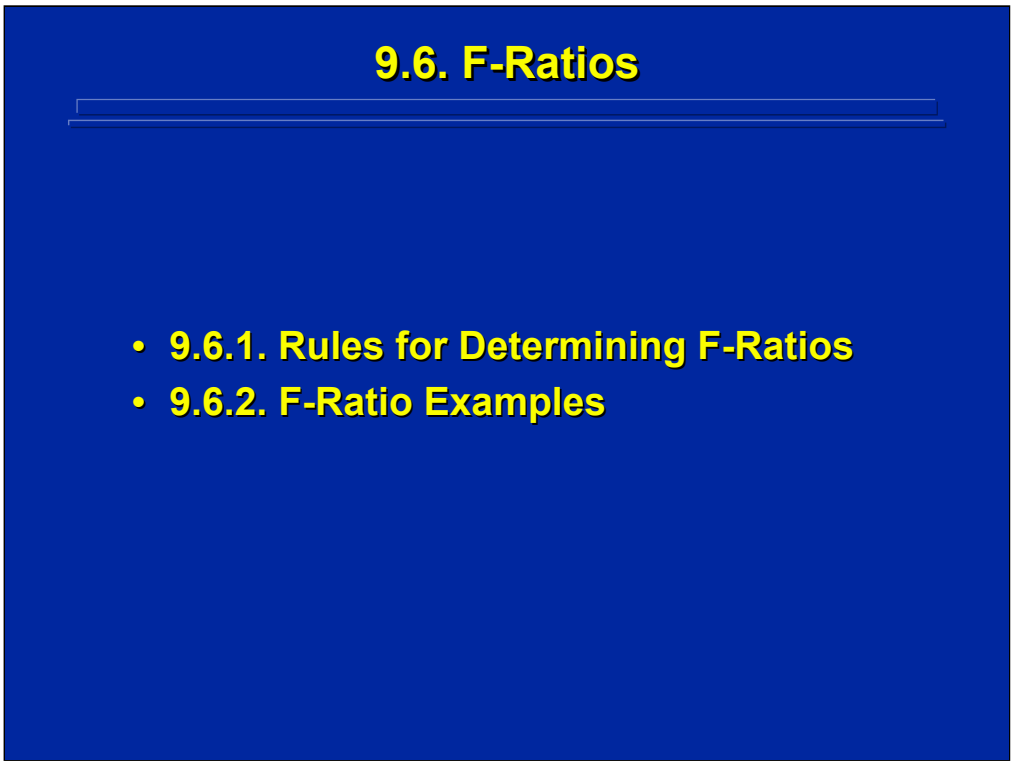

This slide provides a summary of the 5 step algorithm used to cast the E(MS) for each source of variance in the ANOVA Summary Table of this two-factor, between-subjects design in which both factors A and B are fixed-effect factors and the subject effect is a random-effects variable. Remember this algorithm is not a mathematical derivation. Rather, it simply provides rules for stating the resulting E(MS) for any ANOVA design.



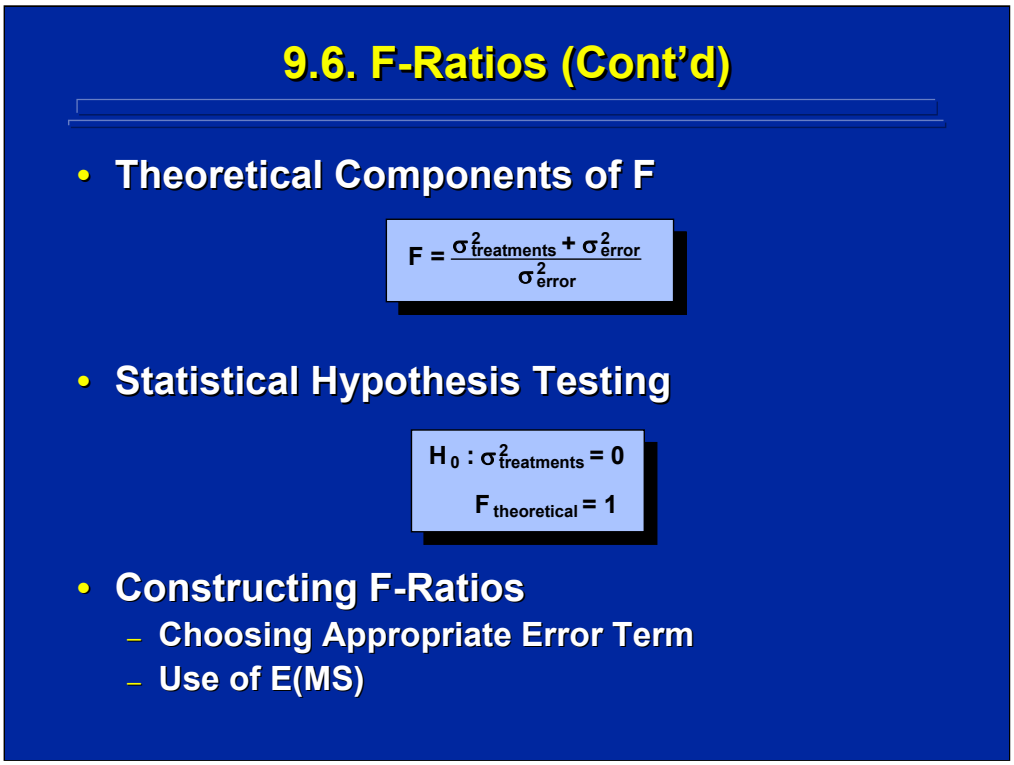

This slide summarizes the results of using the 5 step algorithm to cast the E(MS) of the within-subjects design alternative for a two-factor ANOVA design when factors A and B are both fixed-effects variables and subjects are considered random-effects.




This slide summarizes the results of using the 5 step algorithm to cast the E(MS) of a mixed-factors design alternative for a two-factor ANOVA design where Subjects are nested in Factor A and crossed with Factor B as shown in the statistical model. Again, factors A and B are both fixed-effects variables, and subjects are considered random-effects in casting the E(MS).

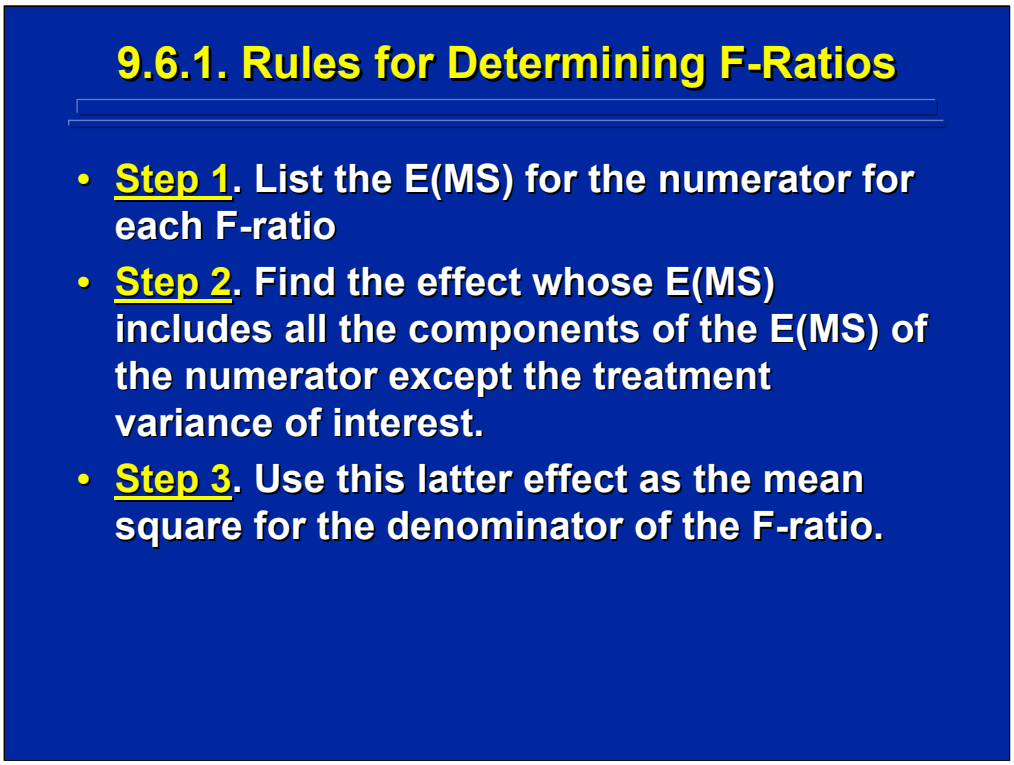



This final piece of the ANOVA Summary Table puzzle is the listing of the possible F-ratios that can be tested in the experimental design. These F-ratios represent the observed values calculated from the results of the experiment that are used in statistical hypothesis testing. In this subsection, rules for constructing these F-ratios are presented and between-subjects, within-subjects, and mixed-factors examples are provided for a two-factor ANOVA design.

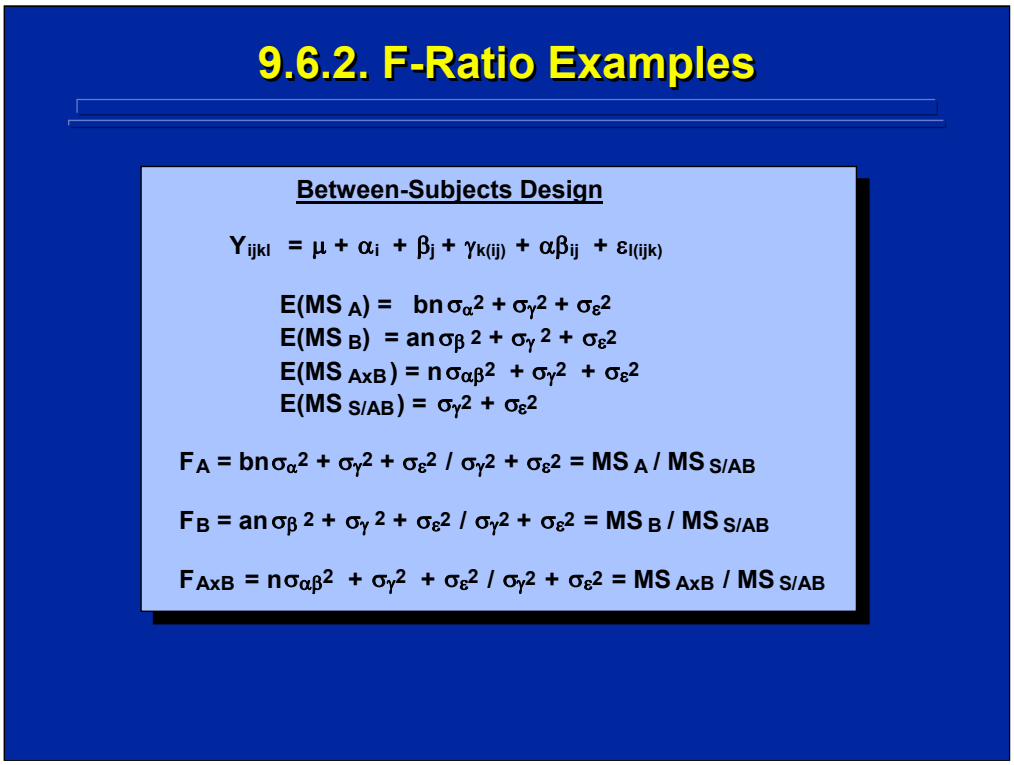



In ANOVA hypothesis testing, the F-ratio is based on two sample variances, and the estimate of the treatment effect is placed in the numerator. As shown on this slide, an F-ratio is theoretically composed of variance due to treatments plus variance due to error components in the numerator and only variance due to error components in the denominator. Recall that the variance due to treatments is 0 under the null hypothesis. Consequently, a theoretical F-value equals 1 when the null hypothesis is true.

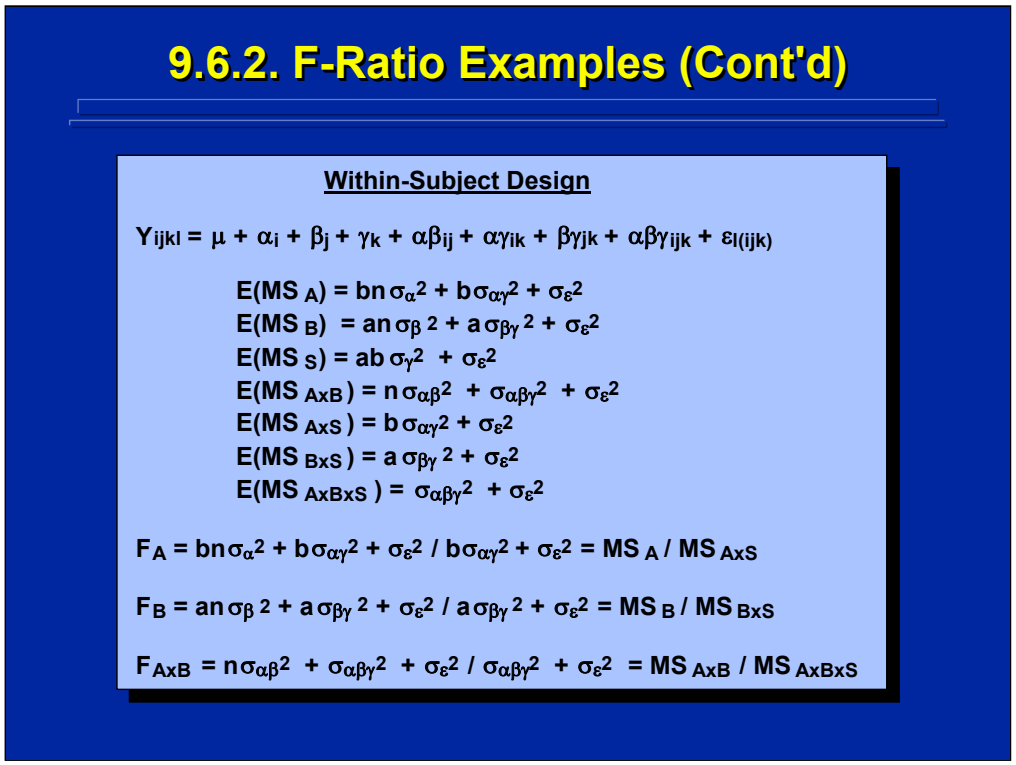
The MS chosen for the numerator is the treatment effect. Then, the MS chosen for the denominator should represent only the error variance components that are theoretically present in the numerator. Consequently, constructing an F-ratio for hypothesis testing is reduced to choosing the appropriate "error term" for the denominator. The choice of the appropriate error term is based on E(MS) because they specify the components of variance present in a particular MS.



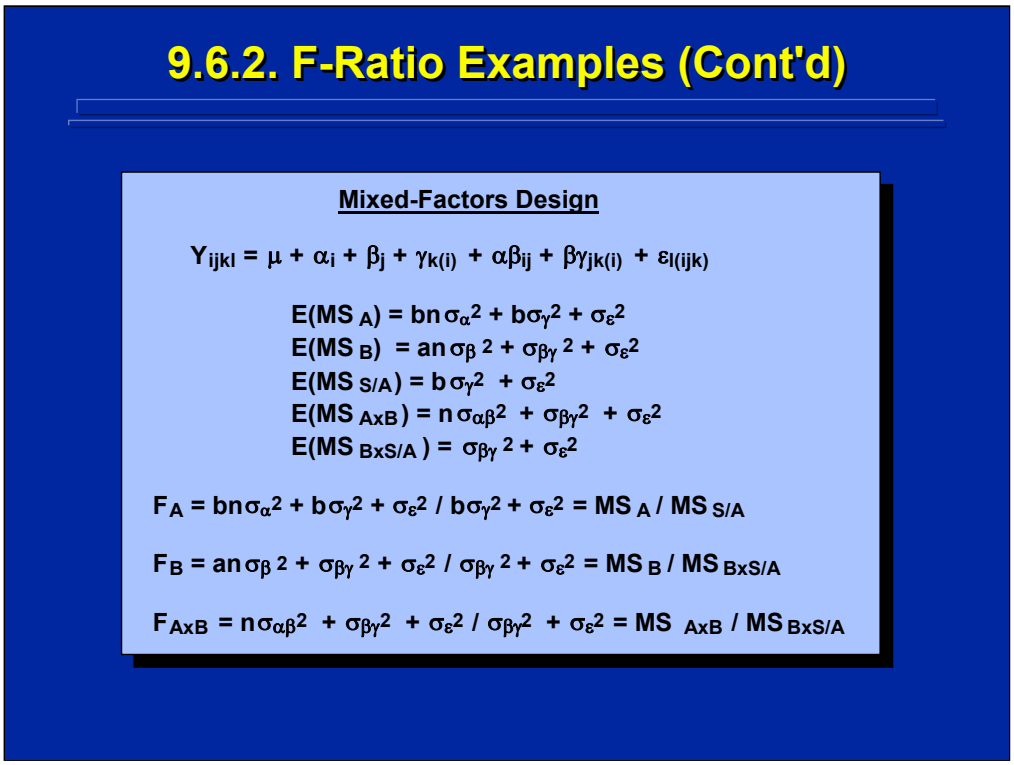

By way of summary, this slide lists three steps for constructing F-ratios based on E(MS). One simply chooses a denominator for the F-ratio where the E(MS) has all the components of the numerator except the treatment effect. If an appropriate denominator, or error term, does not exist for an effect, then that effect cannot be tested using a standard F-test.




This slide illustrates the use of the three steps for constructing F-ratios in a two-factor, between-subjects ANOVA design. Note that only the main effect of Factors A and B as well as the AxB interaction can be tested. In all three hypothesis tests, the E(MS) of the numerator show that $MS_{S/AB}$ should be used as the error term, or denominator, of the F-ratio. In addition, the S/AB effect cannot be tested because no error term exists since none of the E(MS) in the design includes just $\sigma_\varepsilon^2$. Consequently, $MS_{S/AB}$ is only used as an error term in the between-subjects design.

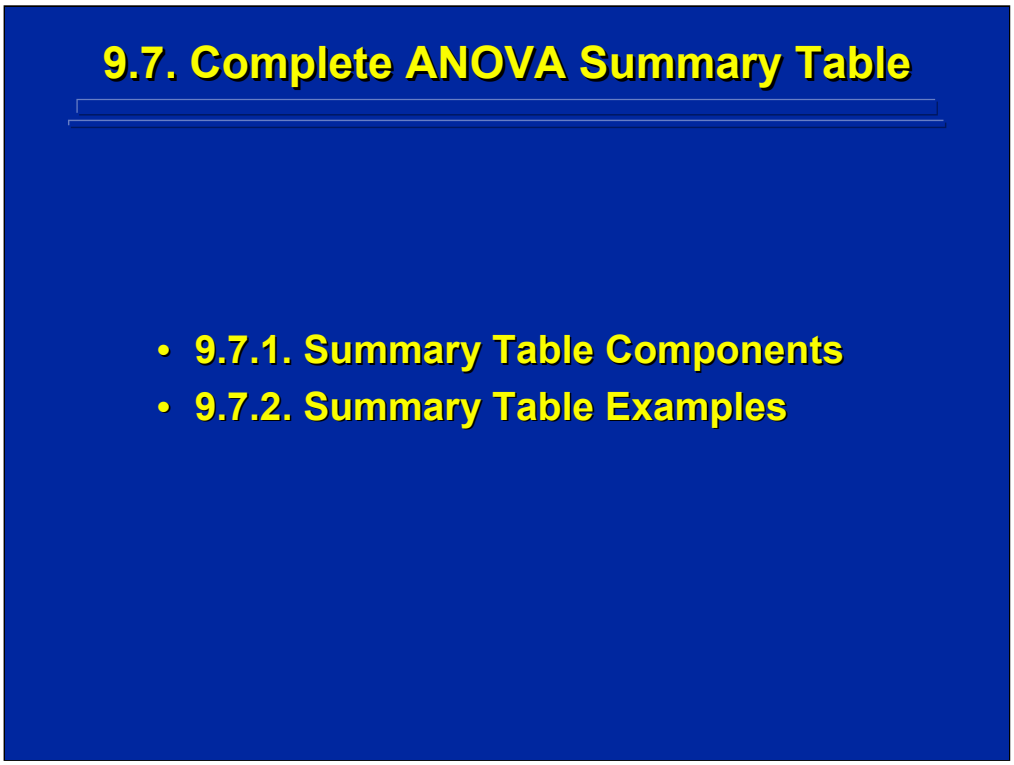



This slide illustrates using the 3 step procedure for constructing F-ratios in a two-factor, within-subjects ANOVA design. Again only the main effect of Factors A and B and the AxB interaction can be tested. In this design, however, the error terms are different. For $F_A$, the appropriate denominator based on E(MS) is $MS_{AxS}$. For $F_B$, the denominator is $MS_{BxS}$. And, for $F_{AxB}$, the denominator is $MS_{AxBxS}$. There are no appropriate F-ratios to test the S, AxS, BxS, and AxBxS effects in the within-subjects design.

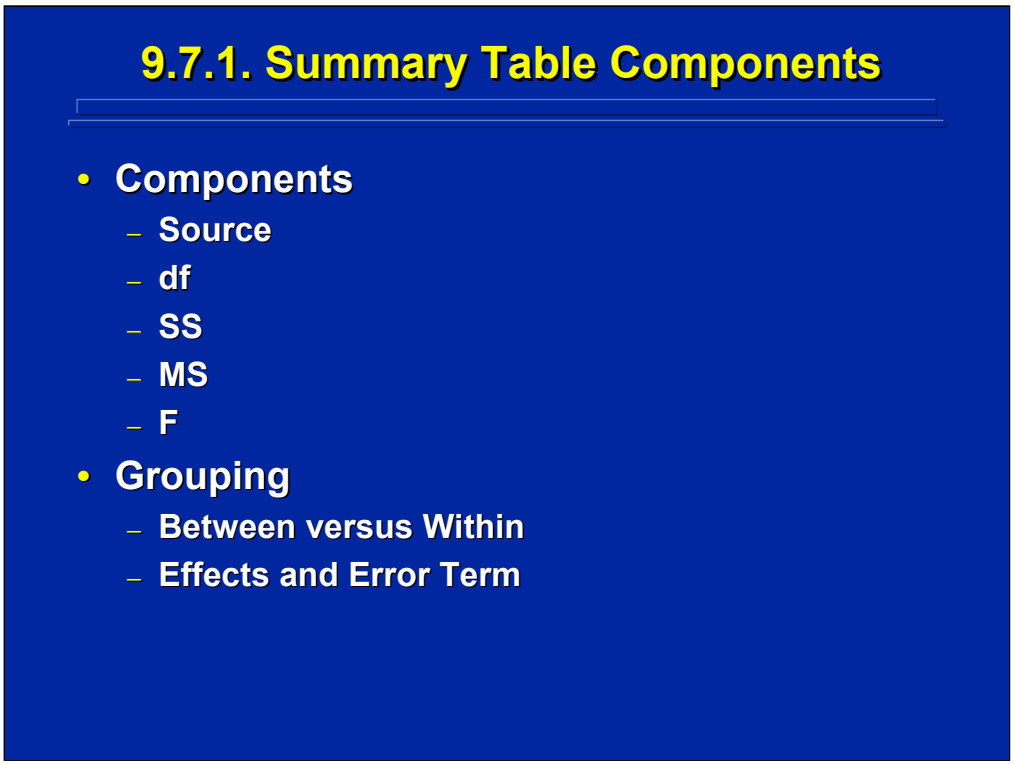



This slide illustrates using the 3 step procedure for constructing F-ratios in a two-way, mixed-factors ANOVA design. Again only the main effect of Factors A and B and the AxB interaction can be tested. In this design, however, the error terms are different. $MS_{S/A}$ is used as the error term in $F_A$. But, $MS_{BxS/A}$ is used as the error term for both $F_B$ and $F_{AxB}$. There are no appropriate error terms to test the S/A or BxS/A effects, and they only exist as error terms in this mixed-factors design.

Whether one collects the data using a two-factor between-subjects, within-subjects, or mixed-factors design, one can always test the A main effect, the B main effect, and the AxB interaction. The only difference between each of these three design alternatives is the error term used in the denominator of the F-ratio. Expected mean squares allow one to determine the appropriate error term in each case.




This subsection puts all the components together into a complete ANOVA Summary Table. First, some general conventions for stating Summary Tables of human factors experiments are presented. Next, examples of the two-factor, between-subjects, within-subjects, and mixed-factors design alternatives are provided.




The results of an ANOVA are presented in a standard Summary Table format. Every ANOVA Summary Table includes five column headings listed in the order shown under the Components portion of this slide. The statistical model provides the row listing of the main effects and interactions included under Source. The degrees of freedom depend upon the number of levels of each factor in the experiment as well as the number of subjects observed in each cell of the design. Before conducting any statistical analysis, the experimenter should list the sources and degrees of freedom of the design in order to facilitate checking the empirical results of an ANOVA conducted with a statistical analysis package.

By convention, the sources listing of rows are grouped into between-subjects and within-subjects effects. In addition, all the effects that use the same error term to form the F-ratio are grouped together with the error term listed last in the grouping. These conventions facilitate reading and checking the ANOVA Summary Table entries. Examples for stating a two-factor ANOVA design in this general format are provided for between-subjects, within-subjects, and mixed-factors design alternatives.

## 9.7.2. Summary Table Examples

**Between-Subjects Design**

| Source | df | SS | MS | F |
|---|---|---|---|---|
| A | a-1 | $SS_A$ | $MS_A$ | $MS_A/MS_{S/AB}$ |
| B | b-1 | $SS_B$ | $MS_B$ | $MS_B/MS_{S/AB}$ |
| AxB | (a-1)(b-1) | $SS_{AXB}$ | $MS_{AXB}$ | $MS_{AxB}/MS_{S/AB}$ |
| S/AB | ab(n-1) | $SS_{S/AB}$ | $MS_{S/AB}$ | |
| Total | abn-1 | $SS_{Total}$ | | |

This slide shows an example of general format of the ANOVA Summary Table for a two-factor, between-subjects design. Notice that the A, B, and AxB effects are listed before S/AB because S/AB is the error term for each of these effects. Since both the df and SS are additive, a total is provided for these two columns. The MS is simply the SS divided by df for each particular source. Finally, the F-ratio is determined by E(MS) in order to determine the appropriate error MS to use in the denominator. The three F-ratios for the between-subjects design all use the same error term, $MS_{S/AB}$. There is no F-ratio for S/AB because no error term exists for this effect.

## 9.7.2. Summary Table Examples (Cont'd)

**Within-Subjects Design**

| Source | df | SS | MS | F |
|---|---|---|---|---|
| **Between** | | | | |
| S | n-1 | $SS_S$ | | |
| **Within** | | | | |
| A | a-1 | $SS_A$ | $MS_A$ | $MS_A/MS_{AxS}$ |
| AxS | (a-1)(n-1) | $SS_{AxS}$ | $MS_{AxS}$ | |
| B | b-1 | $SS_B$ | $MS_B$ | $MS_B/MS_{BxS}$ |
| BxS | (b-1)(n-1) | $SS_{BxS}$ | $MS_{BxS}$ | |
| AxB | (a-1)(b-1) | $SS_{AxB}$ | $MS_{AxB}$ | $MS_{AxB}/MS_{AxBxS}$ |
| AxBxS | (a-1)(b-1)(n-1) | $SS_{AxBxS}$ | $MS_{AxBxS}$ | |
| **Total** | abn-1 | $SS_{Total}$ | | |

This slide shows an example of general format of the ANOVA Summary Table for a two-factor, within-subjects design. Notice that subjects, S, is listed first as a between-subjects effect. Since there is no error term for S, the S effect is not tested, but the df and SS are usually listed to check totals.

All three tested effects in this design are within-subjects effects. Notice that A, B, and AxB are grouped separately with their appropriate error term based on E(MS). So, each of the three F-ratios in the within-subjects design alternative has a different MS denominator that reflects their different error terms.

## 9.7.2. Summary Table Examples (Cont'd)

**Mixed-Factors Design**

| Source | df | SS | MS | F |
|---|---|---|---|---|
| **Between** | | | | |
| A | a-1 | $SS_A$ | $MS_A$ | $MS_A/MS_{S/A}$ |
| S/A | a(n-1) | $SS_{S/A}$ | $MS_{S/A}$ | |
| **Within** | | | | |
| B | b-1 | $SS_B$ | $MS_B$ | $MS_B/MS_{BxS/A}$ |
| AxB | (a-1)(b-1) | $SS_{AxB}$ | $MS_{AxB}$ | $MS_{AxB}/MS_{BxS/A}$ |
| BxS/A | a(b-1)(n-1) | $SS_{BxS/A}$ | $MS_{BxS/A}$ | |
| **Total** | abn-1 | $SS_{Total}$ | | |

This slide shows an example of general format of the ANOVA Summary Table for a two-factor, mixed-factors design. Notice that both between-subjects and within-subjects effects are tested in this design alternative. One can easily tell that A is the between-subjects factor and B is the within-subjects factor in this design. Factor A is grouped with its error term, S/A, as between-subjects effects, and both B and AxB are grouped with their error term, BxS/A, as within-subjects effects.

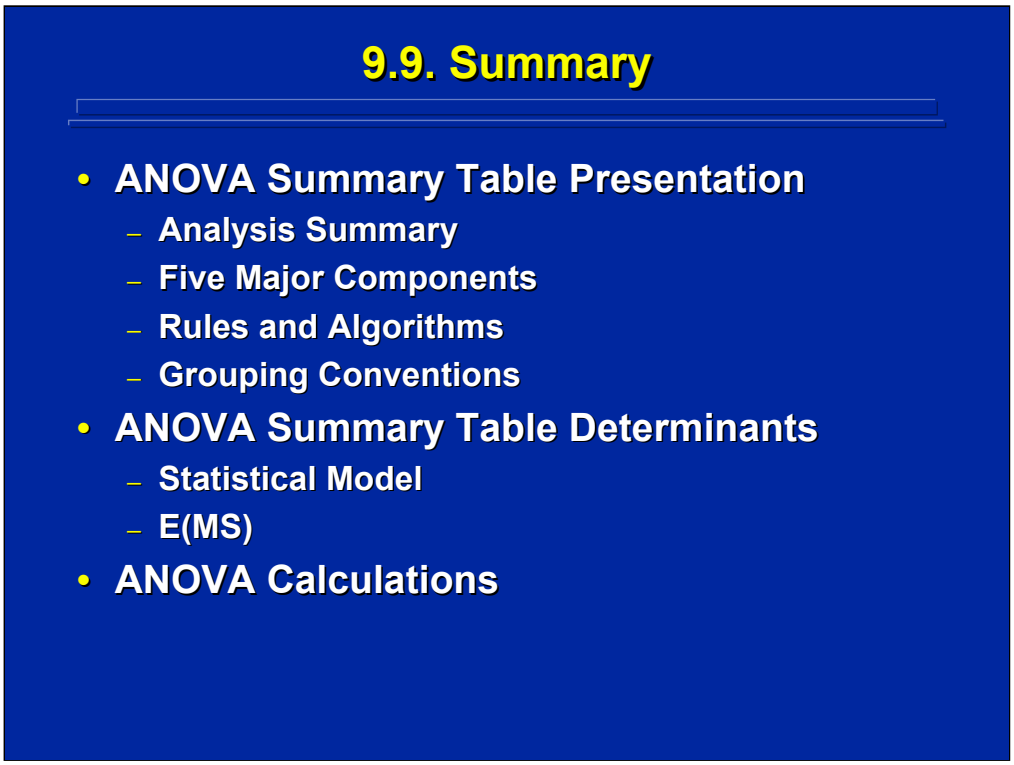
Notice that A, B, and AxB are the only three effects that can be tested in either the between-subjects, within-subjects, or mixed-factors alternative of this two-factor ANOVA design. However, the three alternatives differ in terms of the error terms used in the F-ratios. The grouping conventions for specifying ANOVA Summary Tables help to highlight and check these differences.

## 9.8. ANOVA Design Construction

- Determine the Number of Factors
- Determine the Levels of Each Factor
- Specify the Type of Factor (Fixed or Random)
- Specify the Relationship of the Factors (Crossed or Nested)
- Classify the Design (Between, Within, or Mixed)
- State the Statistical Model
- List E(MS)
- Determine the F Ratio
- List the ANOVA Summary Table

The major steps involved in constructing any ANOVA experimental design are listed on this slide in order of consideration. The number of factors and levels of each factor determine the configuration of the factorial design. The relationship of the factors determine the design classification, statistical model, and E(MS), possible F-ratios. Once these steps are completed, the experimenter can specify the general format of the ANOVA Summary Table.

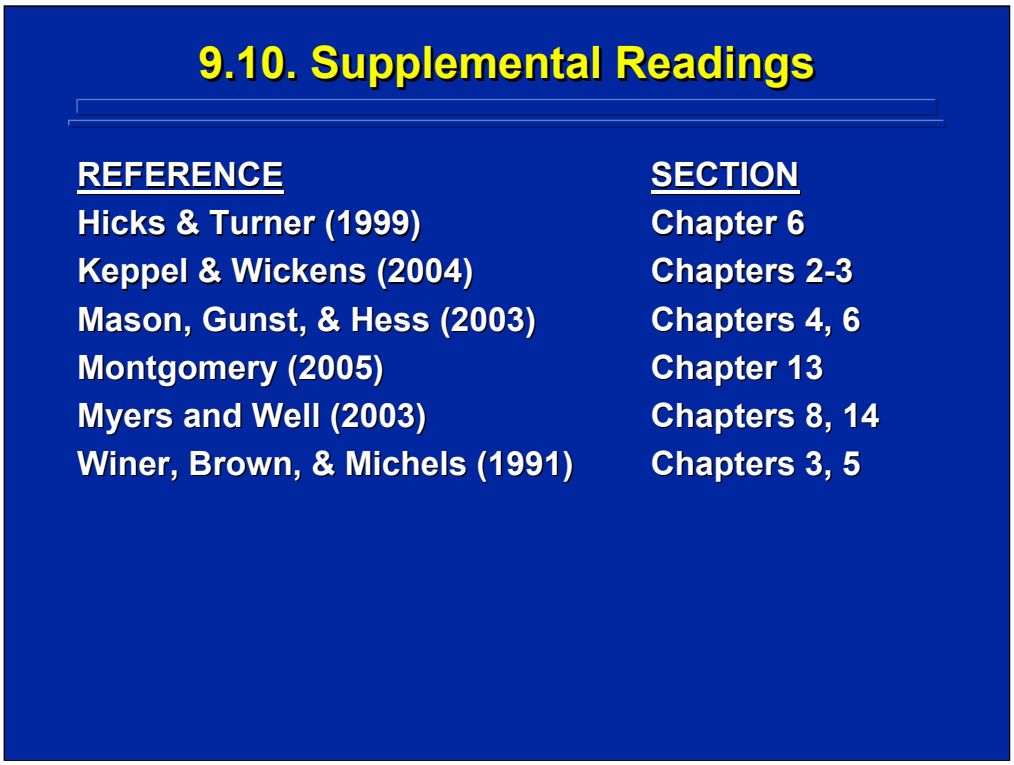
All of these tasks need to be completed before collecting any data in an experiment to be sure that the planned design will provide the necessary data to evaluate the hypotheses of interest in the research. The general format of the ANOVA Summary Table allows for straightforward checks of subsequent numerical output of statistical analyses of the data collected in the experiment.




A Summary Table is the standard way of presenting the results of ANOVA calculations. The five major components of the Summary Table include the Sources, df, SS, MS, and F-ratios. Rules and algorithms rather than derivations were presented as a means of facilitating specification and calculation of the components. Standard conventions were discussed for grouping and presenting sources in ANOVA Summary Tables used in human factors research.

The statistical model of an ANOVA design is the major determinant of the Summary Table because it specifies all the sources of variance that can be calculated and defines the crossed and nested relationships among factors that determine between-subjects, within-subjects, and mixed-factors designs. The researcher must understand the concept of E(MS) in order to determine the appropriate error term to use in the denominator of F-ratios for various design alternatives.

The experimenter should list the sources and degrees of freedom of the design before calculating an ANOVA as a simple check for computation errors. An algorithm for generating SS formulae will be described in Topic 9 that the researcher can use to make all the numerical calculations in an ANOVA Summary Table. However, statistical analysis packages are usually used for these computations in most experiments. The use of SAS for conducting ANOVA analyses is described in Slater and Williges (2006).



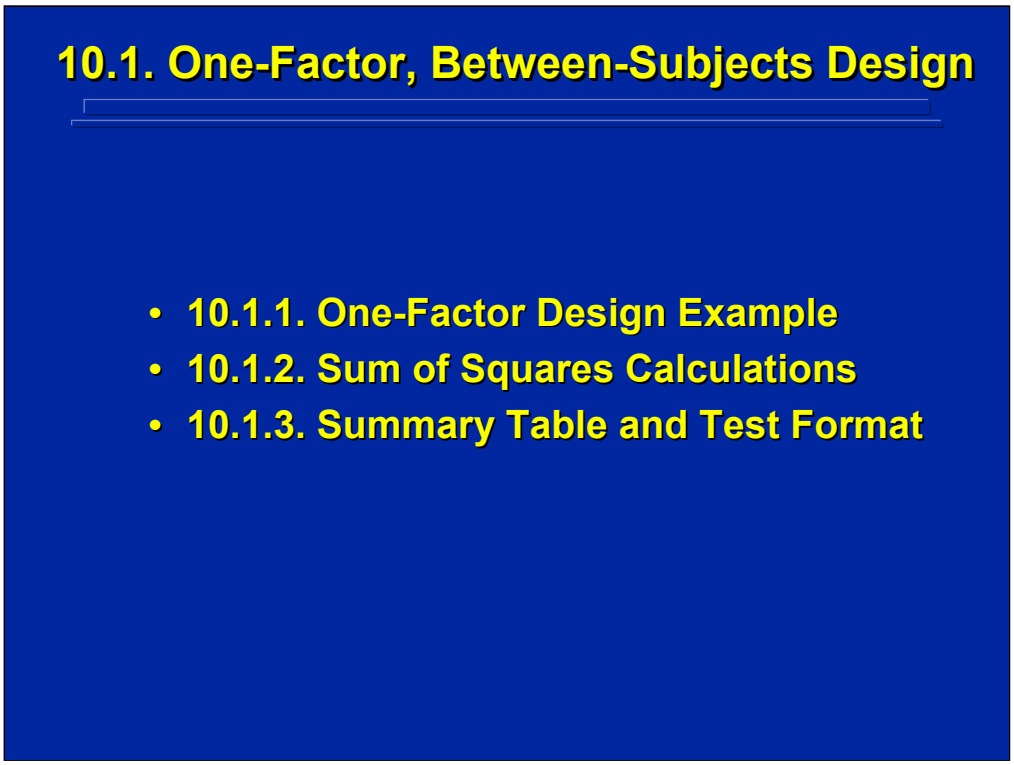

Both Keppel and Wickens (2004) and Winer, Brown, and Michels (1991) provide general descriptions of various components of the ANOVA Summary Table in detail. In addition, all six references listed on this slide describe various rules and algorithms for determining df and E(MS).

# Topic 10. Between-Subjects ANOVA Designs

10.1. One-Factor, Between-Subjects Design
10.1.1. One-Factor Design Example
10.1.2. Sum of Squares Calculations
10.1.3. Summary Table and Test Format
10.2. Two-Factor, Between-Subjects Design
10.2.1. Design Configuration
10.2.2. AxB Interaction
10.2.3. Calculations
10.2.4. Two-Factor Design Example
10.3. n-Factor, Between-Subjects Design
10.4. Summary
10.5. Supplemental Readings

This topic covers the construction and computational details of the first of the three major categories of ANOVA designs used in human factors and ergonomics research. Between-subjects designs are discussed in terms of one-factor, two-factor, and n-factor designs. The procedures for determining sum of squares computational formulae for any ANOVA are presented in the discussion of one-factor designs. The concept of an interaction is presented in two-factor designs. Generalizations for constructing and analyzing any between-subjects ANOVA design are summarized under n-factor designs. Computational examples are provided for both a one-way and a two-way between-subjects design.

A summary listing of all the ANOVA procedural rules and algorithms for conducting any ANOVA analysis in human factors research is provided at the end of this topic. References to supplemental readings on between-subjects designs are provided for the major experimental design texts appropriate for human factors research.




Between-subjects designs use a different group of randomly assigned subjects in each treatment combination. The treatments in these completely randomized designs can consist of any number of factors and any number of levels of each factor. The simplest between-subjects design has only one factor with two-levels. As shown in Topic 8, the analysis of this simple design reduces to a standard t-test of two means.

In order to describe the SS computations in ANOVA, this subsection uses a one-factor, between-subjects design with three levels in which the overall difference among three treatment means is assessed. If a significant difference is found in the F-test, at least one of the paired differences among the three treatment means is significant. A general algorithm for generating SS computational formulae and a numerical example using these computational formulae are provided.

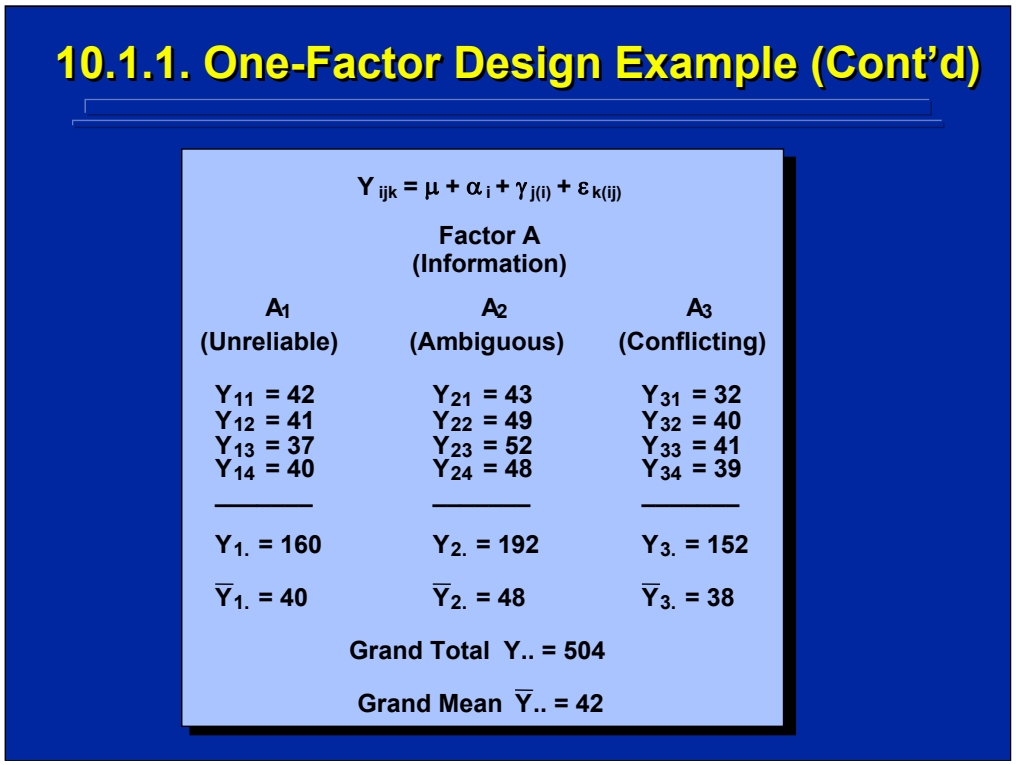
## 10.1.1. One-Factor Design Example

- **Example Problem: The effect of various aspects of information in military command and control situations was evaluated in terms of a commander's situation awareness. Situation awareness was measured for each of four different commanders who received information characterized as unreliable, ambiguous, or conflicting. Each commander received only one of the three types of information. Do these three aspects of information have a significant effect on a commander's situation awareness ($p < 0.05$)?**

(Click in this red rectangle to see SAS calculations for this example.)

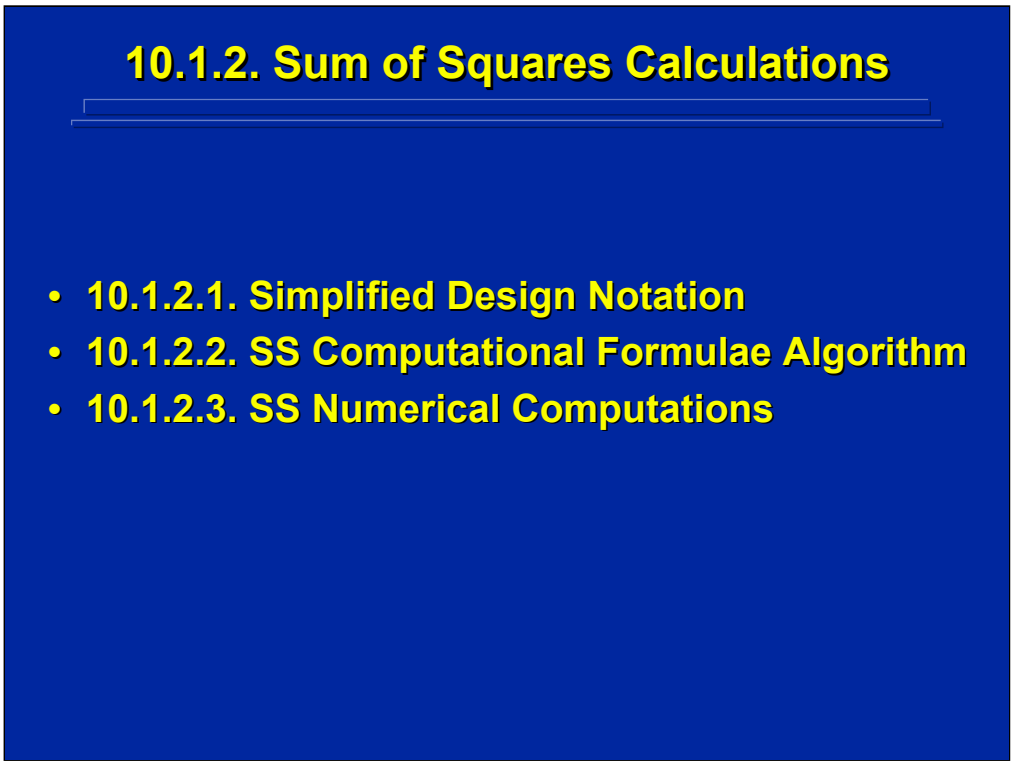
This example problem is a between-subjects design because 4 different commanders are used in each treatment condition resulting in 12 different subjects in the complete experiment. The one factor, information, investigated in this experiment has three levels: unreliable, ambiguous, and conflicting. Hence the experiment to evaluate level of a commander's situation awareness uses essentially a one-factor, three-level, between-subjects design.

This reference material demonstrates the hand calculations for conducting the ANOVA on data obtained from this example problem. Due to the effort involved in calculating the SS in complex ANOVA design, most human factors researchers use statistical analysis packages for conducting ANOVAs to facilitate analysis effort and reduce computational errors. Slater and Williges (2006) appendix provides the results of this ANOVA using the SAS computer package.



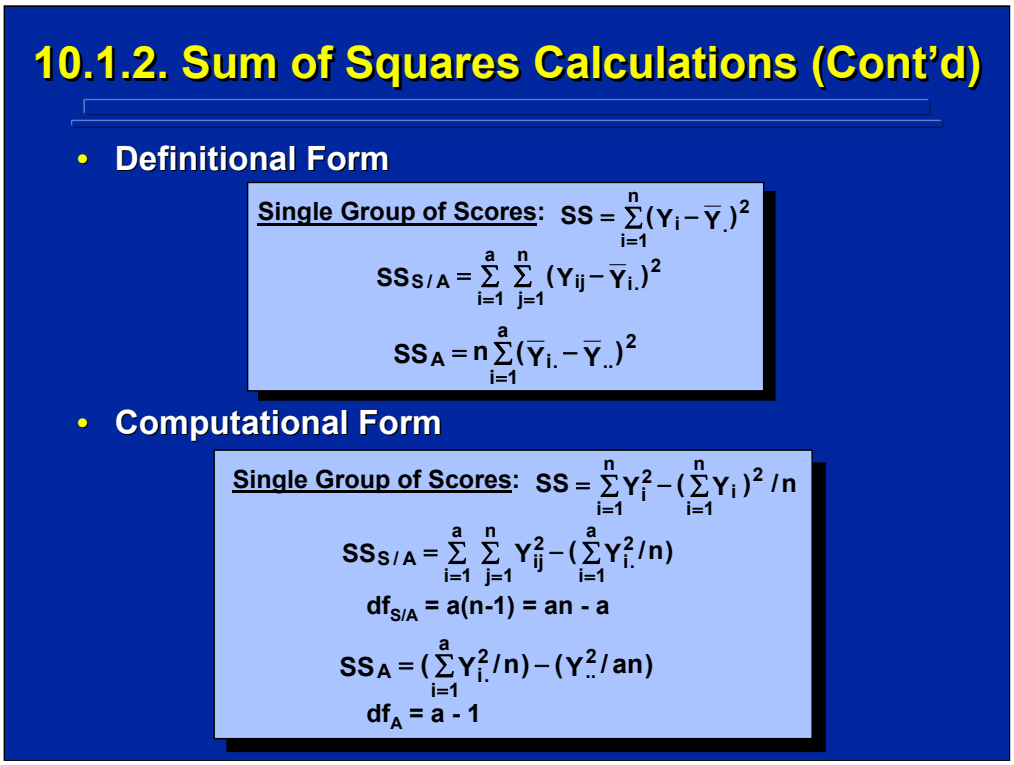

(Click in this red rectangle to see SAS calculations for this example.)

Hypothetical data obtained from the situational awareness experiment are shown on this slide in the dotted notation to indicate sums and bars for mean values. The statistical model for the one-factor, between-subjects design is shown on the top of the slide. Note that the actual name of Factor A and the names of the 3 levels of the Factor A are listed in parenthesis on this slide.




(Click in this red rectangle to see SAS calculations for this example.)

In order to conduct the ANOVA on the example data, the experimenter must calculate the SS associated with each of the Sources of Variation. This subsection describes a simplified design notation that facilitates stating computational formulae for complex ANOVA designs. An algorithm is provided for generating SS computational formulae based on this simplified notation, and calculations of the SS for the one-way, between-subjects example problem are presented to demonstrate use of this algorithm.

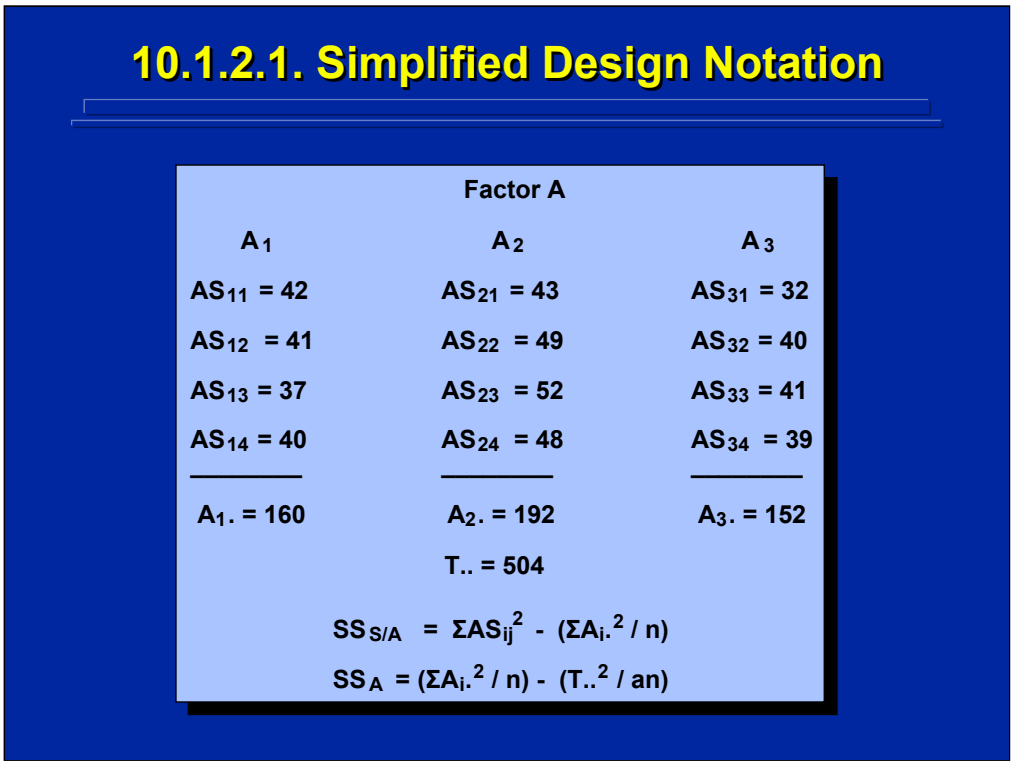



(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the general formula for both the definitional and computational forms for the sum of the squared deviations around the mean for a single group of scores showing the index of summation, n, (Myers, 1979, pp. 15-16). In addition, the formulae for calculating the $SS_{S/A}$ and $SS_A$ in the two-group example problem are also provided in both forms on this slide. Myers (1979, pp. 76-83) shows the derivation of the computational form from the definitional form and demonstrates the isometric relationship between the summation indexes "a" and "n" of the SS computational components and the expanded df of S/A and A.

Although both forms of the SS formulae are algebraically equivalent, note that the definitional form includes means, and the computational form includes only sums. To avoid rounding errors and to enhance ease of calculation, the computational form is usually preferred.

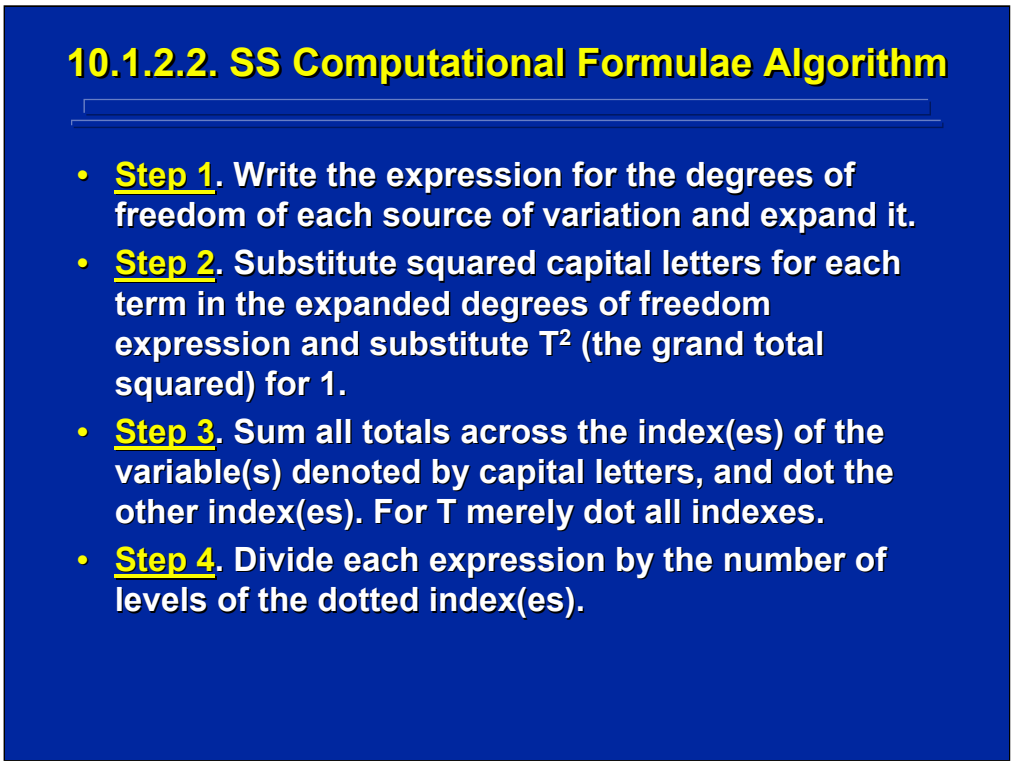
Since the scores in the one-factor example problem represent different levels of Factor A and different subjects, S, double summation signs are required with indexes "a" and "n", respectively, to specify the sum of individual observations in the experiment using standard Y notation as described in Chapter 2 of Myers (1979). As the number of factors increases in complex ANOVA designs, the number of summation signs also increases in the SS formulae, and the standard Y notation becomes more cumbersome.




(Click in this red rectangle to see SAS calculations for this example.)

This slide shows a simplified design notation similar to the notation used by Keppel and Wickens (2004, pp.26-30) that avoids use of multiple summation signs required in the standard Y notation. Each individual Y observation is specified in terms of the effects represented by the subscripts. In the one-factor example problem, each observation represents a particular level of Factor A and a particular subject. Rather than designate the observation as Y, it is designated as AS with the appropriate subscripts. Totals for each level of Factor A are designated just by the particular level of A and the subscript for S is dotted to designate summing across all levels of S. The grand total of all scores is represented by T and dotted across all subscripts. This notation can be extended to multifactor designs by simply adding more letters to the individual observation to represent each additional factor. Various group totals are represented by the appropriate letters and dotted subscripts.

The data matrix for the example problem is restated using this simplified notation. In addition, the SS computational formulae can all be stated with the use of a single summation sign as shown on the bottom portion of this slide. The summation sign designates the sum of the squared raw score or group total represented by the simplified notation. All ANOVA designs and computational formulae referred to in this reference material will use this simplified notation.




(Click in this red rectangle to see SAS calculations for this example.)

This slide shows a 4 step algorithm patterned after Keppel and Wickens (2004, pp. 216-218) for generating SS computational formulae using the simplified design notation. This algorithm is based on the demonstrated isomorphic relationship of SS formulae components to degrees of freedom. The next two slides provide an example of using each of these four steps in specifying the SS computational formulae for the one-factor, between subjects design.

## 10.1.2.2. SS Computational Formulae Algorithm

- **Step 1. Write the expression for the degrees of freedom of each source of variation and expand it.**

| | A | S/A | Total |
|---|---|---|---|
| Step 1. | a - 1 | an - a | an - 1 |

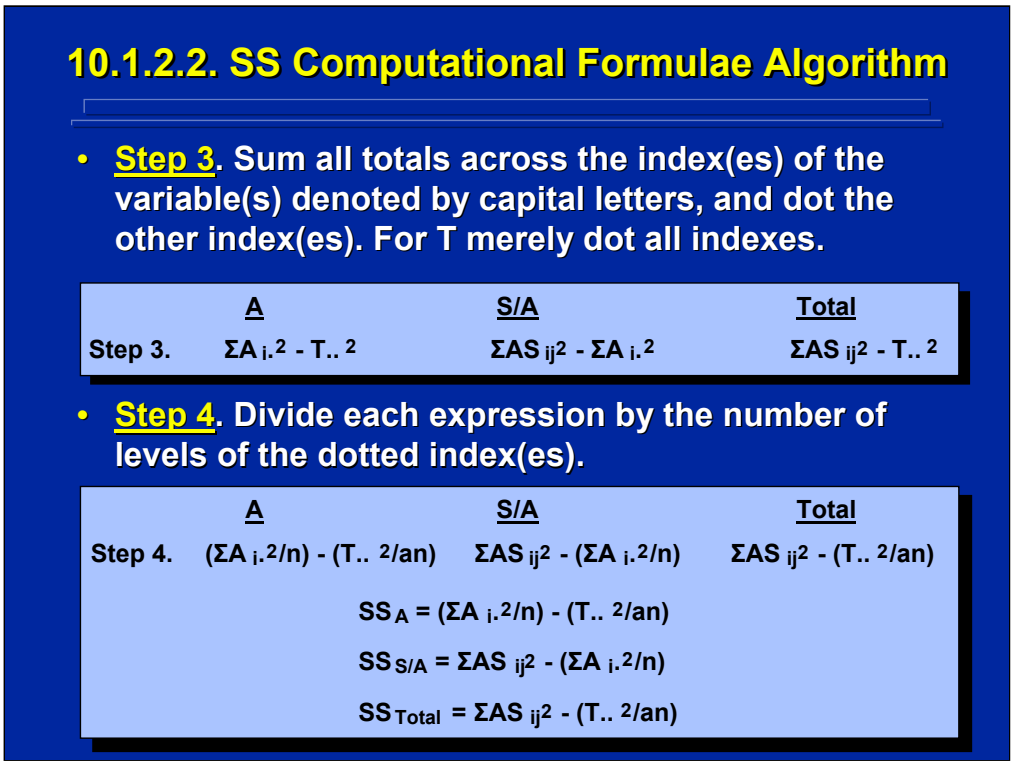
- **Step 2. Substitute squared capital letters for each term in the expanded degrees of freedom expression and substitute $T^2$ (the grand total squared) for 1.**

| | A | S/A | Total |
|---|---|---|---|
| Step 2. | $A^2 - T^2$ | $AS^2 - A^2$ | $AS^2 - T^2$ |

(Click in this red rectangle to see SAS calculations for this example.)

Step 1 is a simple expansion of the degrees of freedom for each source. Remember that “n” is equal to the number of observations in a cell, assuming equal sample size, and not necessarily the number of different subjects appearing in the experiment. For example, “n” in a between-subjects design refers to a different group of “n” subjects in each cell. Likewise, “n” refers to the same subjects that appear in every cell of a within-subjects design, and “n” refers to a combination of the same and different subjects that appear in a mixed-factors design.

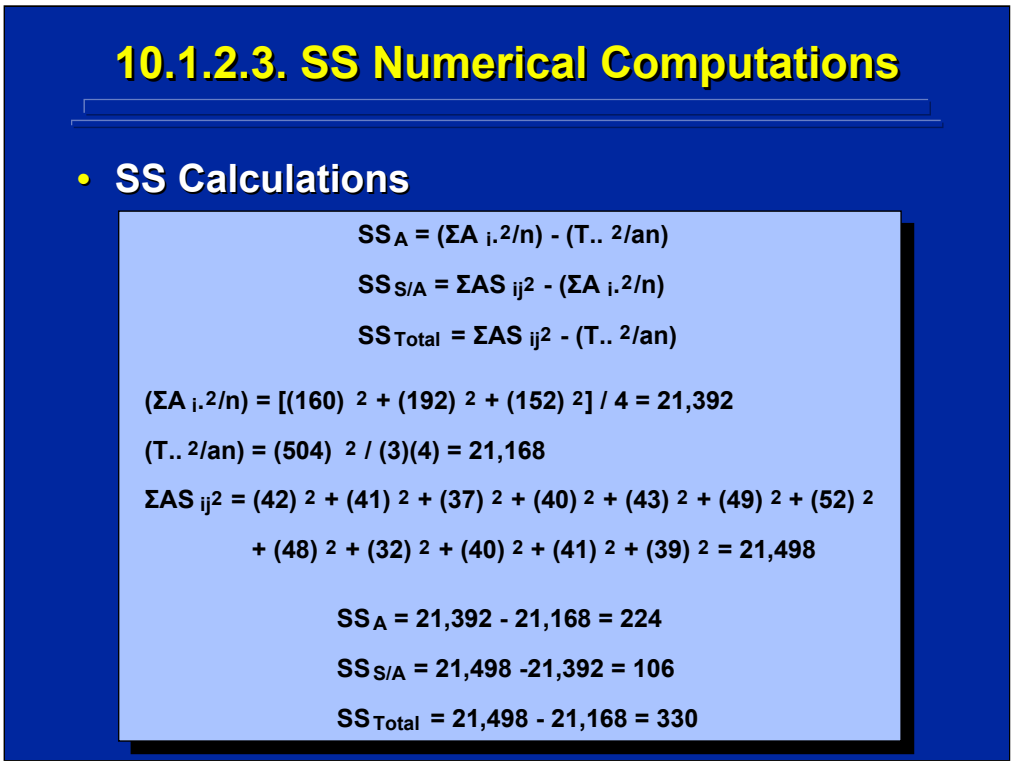
Step 2 substitutes capital letters used in the simplified design notation for the lowercase letters shown in Step 1. Note that the grand total of all scores, T, is substituted for 1 wherever is appears in Step 1. Each of the resulting capital letter combinations is squared.




(Click in this red rectangle to see SAS calculations for this example.)

Step 3 adds the subscripts to the squared values resulting from Step 2 shown on the previous slide. The appropriate subscript starting with "i" is added for each capital letter combination and dots are provided for each letter not included in the combination.

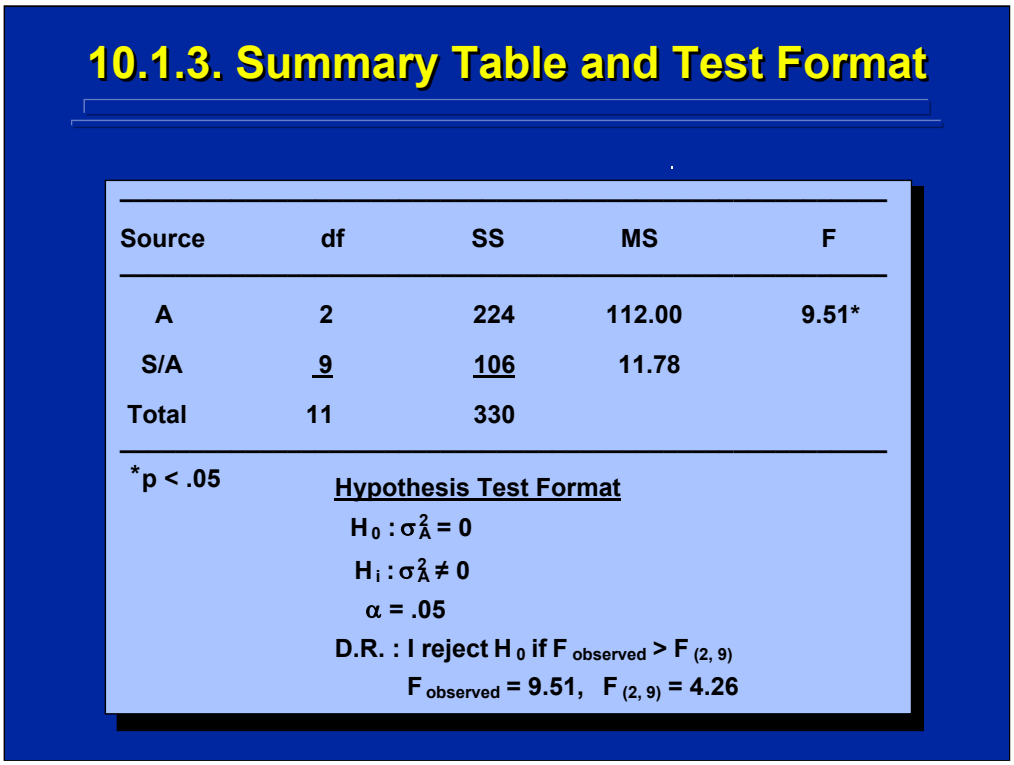
Step 4 simply divides the letter combination in Step 3 by the levels of each dotted index. This provides the final SS computational formula for each source of variation in the design in the simplified design notation. The three SS computational formulae for the one-factor, between-subjects design is shown on the bottom portion of this slide. Note the $AS_{ij}^2$ designates the square of each raw score in the one-way design. These raw scores are squared and then summed. The other letter combinations that are written in parenthesis to designate various squared totals that are summed and then divided by appropriate weights.




(Click in this red rectangle to see SAS calculations for this example.)

Actual SS calculations for the one-factor example problem data are shown on this slide. Note that only three possible components are used in each of the three SS formulae shown on the top portion of the slide. The center portion of the slide shows the calculation of each of these three components based on the simplified notation data matrix for the example problem. The final SS values for the three sources in the example problem are shown on the bottom of this slide.

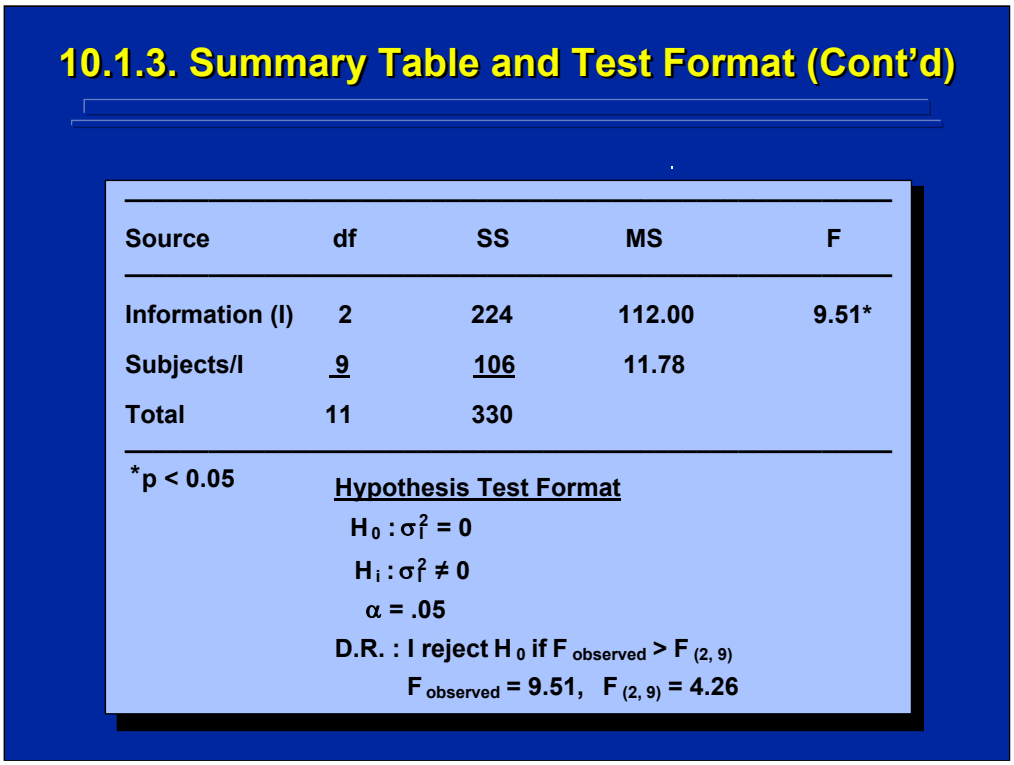
Remember that sum of squares are additive so that $SS_A$ plus $SS_{S/A}$ equal $SS_{Total}$ (i.e., 224+106 = 330). Always calculate $SS_{Total}$ separately and compare it to the sum of $SS_A$ and $SS_{S/A}$ as an easy check for possible calculation errors. If the totals are not equal either some sources(s) in the design were not included or there is an error in calculating one or more of the SS sources. If any of the SS calculations result in a negative number there is an error, because SS values are always positive by definition.




(Click in this red rectangle to see SAS calculations for this example.)

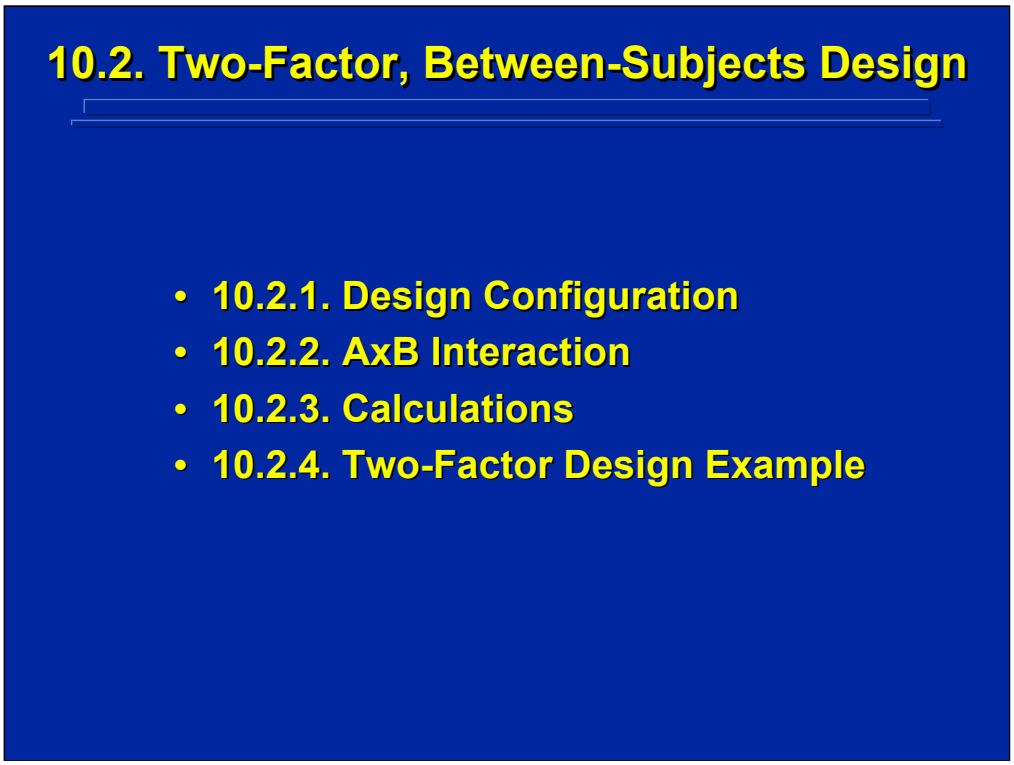
The top portion of this slide shows the final Summary Table for the ANOVA conducted on the example problem data from the one-factor, between-subjects design. The SS are divided by their degrees of freedom to obtain the MS value. The F-ratio for testing Factor A is calculated by dividing $MS_A$ by $MS_{S/A}$ as specified by E(MS).

The standard format for testing the significant difference among the three means is presented in the bottom portion of this slide. Since $F_{Observed}$ is greater than $F_{Tabled}$ there is a significant difference between means at the 0.05 level of significance. This level of significance is noted by the asterisk value in the ANOVA Summary Table. This hypothesis test determines that the main effect of Factor A is significant which means that at least one pair of means is significantly different. Since there are 3 levels of Factor A, there are 3 possible paired differences between means. The overall F-test on Factor A does not specify which paired differences are significant, and additional post hoc analyses are needed to isolate these differences. Procedures for conducting these post hoc tests are presented in Topic 10. From the overall F-test, the experimenter only knows that at least the largest paired difference is significant.

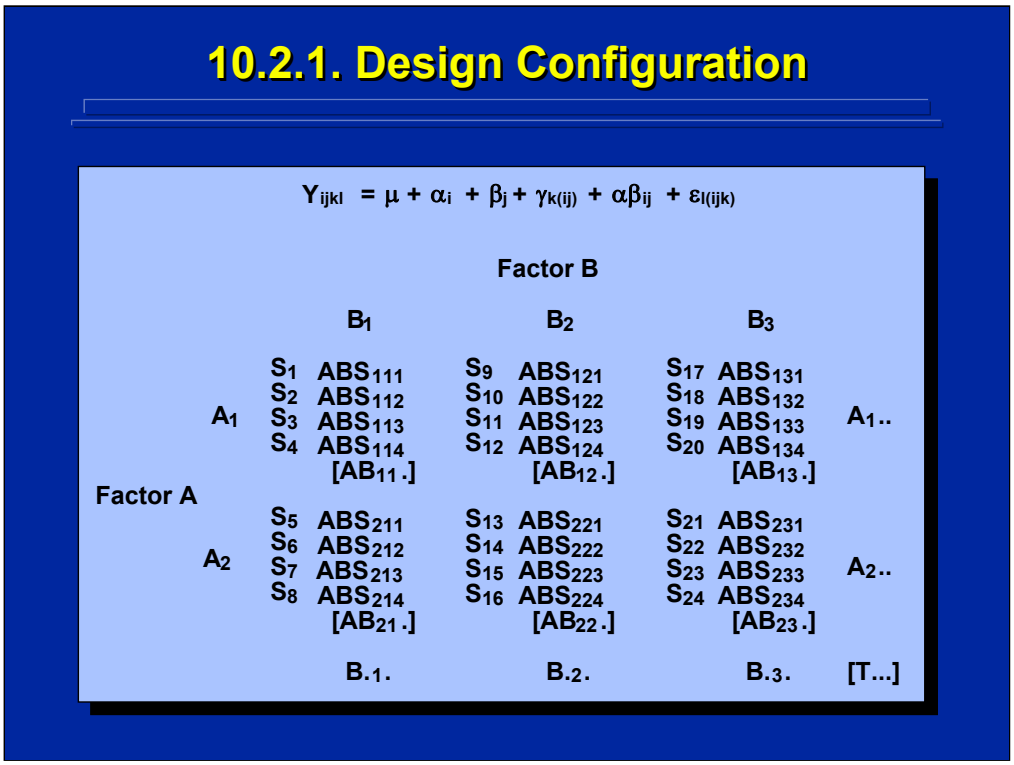



(Click in this red rectangle to see SAS calculations for this example.)

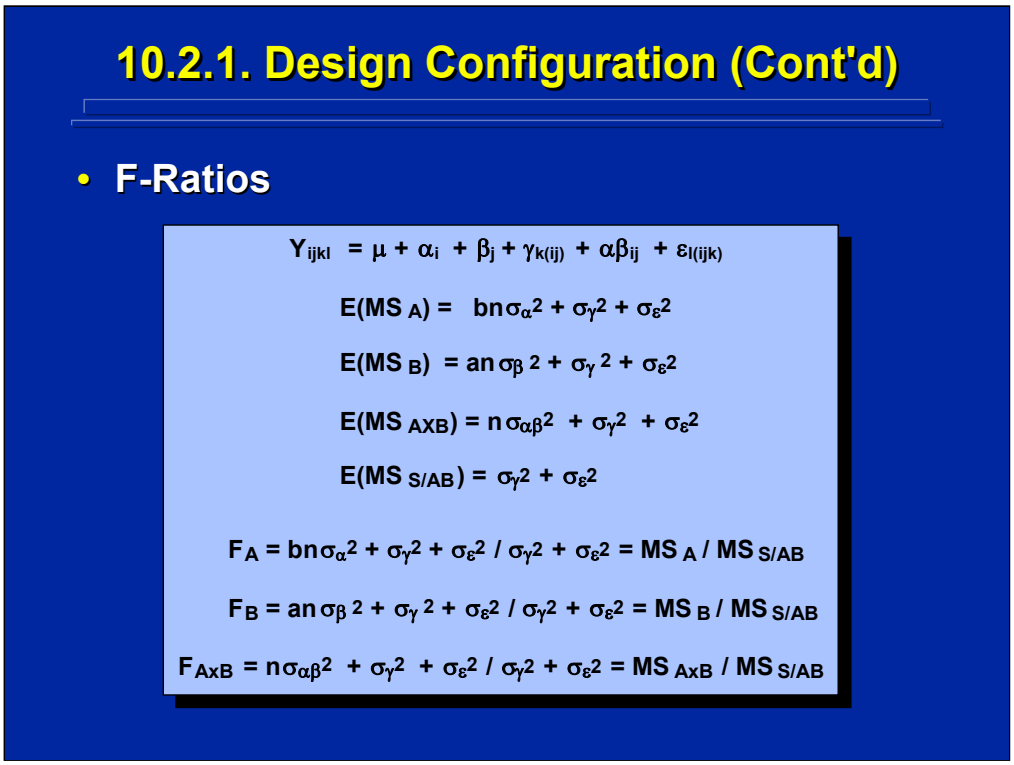
This slide simply restates the previous slide in the context of the example problem related to a commander's spatial ability. Rather than use Factor A, it is more meaningful to list a short, real factor name and a unique one-letter abbreviation, if possible. Hence Factor A is listed as Information (I) in the Summary Table. Based on the results of this ANOVA, the experimenter can conclude that characteristics of information (i.e., unreliable, ambiguous, and confusing) did have a significant effect on a commander's mean level of spatial ability ($p < 0.05$).



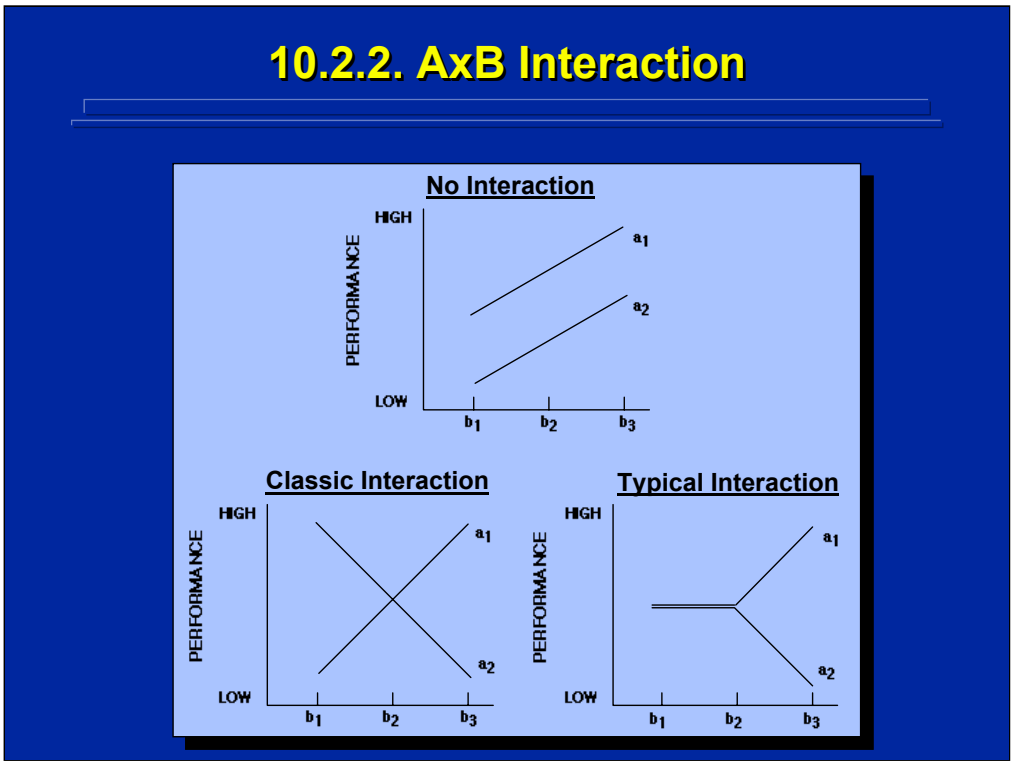

This subsection extends the one-factor design discussion to calculating a two-factor, between-subjects ANOVA. The simplified design notation is extended to include both Factors A and B, and the concept of an interaction between A and B is described. All appropriate computational formulae for this two-factor design are specified using the SS algorithm. Finally, an example problem of a two-factor, between-subjects ANOVA is presented.




This slide shows the general form of a 2x3, two-factor design in the simplified design notation. Note that each individual observation in a two-factor design is designated by ABS. The first subscript is for the level of A, the second is for the level of B, and the third is the level of subjects. Since subjects are nested within Factors A and B in the between-subjects design, there are 24 levels of different subjects as shown on the slide. The simplified notation, however, just designates the 4 subjects in each cell as the subscripts for S. The totals for each level of Factor A are designated by $A_{i..}$, the totals for Factor B are designated by $B_{.j.}$, the totals in each of the 6 cells of the 2x3 design are designated by $AB_{ij.}$, and the grand total of all the scores is designated by $T_{...}$.



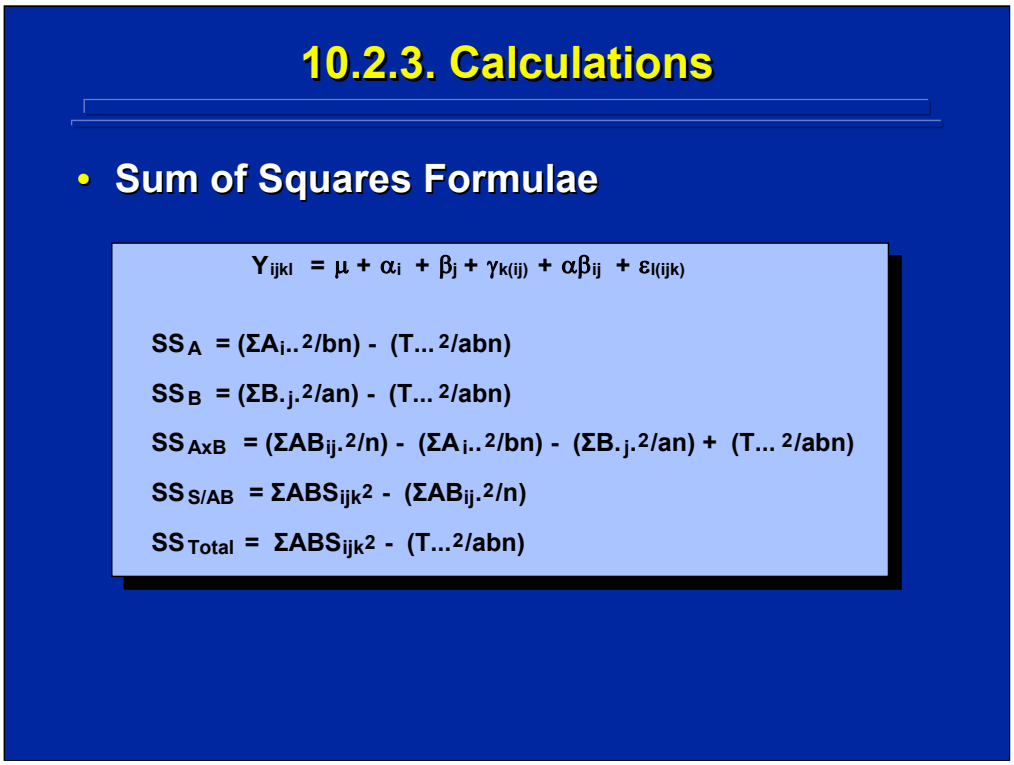

The top of slide states the statistical model for the two-factor, between-subjects design showing Subjects, $\gamma$, nested in both Factors A and B. The expected mean square based on the E(MS) algorithm assuming A and B are fixed-effects variables and S is a random-effects variable are listed in the center portion of this slide. Based on the E(MS) and the rules for constructing F-ratios, the three possible F-ratios for this two-factor design are listed at the bottom of this slide. Note that $MS_{S/AB}$ is the error term in all three F-tests. In summary, the two-factor ANOVA design allows the experimenter to test the main effects of Factors A and B and the AxB interaction.




A factorial experiment allows the experimenter to assess main effects and interactions independently. Recall that an interaction is the differential effect of one factor on another factor. This slide shows stylized plots of hypothetical data resulting in either no interaction, a classic interaction, or the more typical interaction. No interaction exists when the outcomes across levels of one factor are identical at each level of the other factor as depicted in the plot of parallel lines on the top of this slide.

The classic “X” interaction between A and B is shown on the bottom left plot where $a_1$ is less than $a_2$ at level $b_1$, there is no difference at $b_2$, and $a_1$ is greater than $a_2$ at $b_3$. The more typical interaction depicted in the bottom right plot shows that $a_1$ and $a_2$ are only different at $b_3$.

Note that tests of main effects and interactions are independent. For example, the no interaction plot shows differences in both main effects, the classic interaction plot depicts no main effects, and the typical interaction effect shows a possible main effect of only factor A.




The SS computational formulae for the 2x3 between-subjects design based on the SS algorithm are listed on this slide. Note that these formulae are based on only five different computational components.

## 10.2.3. Calculations (Cont'd)

- **ANOVA Summary Table**

| Source | df | SS | MS | F |
|---|---|---|---|---|
| A | a-1 | $SS_A$ | $MS_A$ | $MS_A/MS_{S/AB}$ |
| B | b-1 | $SS_B$ | $MS_B$ | $MS_B/MS_{S/AB}$ |
| AxB | (a-1)(b-1) | $SS_{AxB}$ | $MS_{AxB}$ | $MS_{AxB}/MS_{S/AB}$ |
| S/AB | ab(n-1) | $SS_{S/AB}$ | $MS_{S/AB}$ | |
| Total | abn-1 | $SS_{total}$ | | |

The general form of the two-factor, between-subjects design ANOVA Summary Table is presented on this slide using standard conventions. Note that the error term, S/AB, is listed below A, B, and AxB.

## 10.2.4. Two-Factor Design Example

- **Example Problem: Readability of printed text on a computer screen was evaluated in terms of two fonts (Helvetica and Old English) and number of words displayed per line (10, 20, or 30 words per line). Four different subjects read one particular combination of these two factors, and reading comprehension was tested. Did either of these two factors or the interaction between them have a significant effect on reading comprehension ($p < 0.01$)?**

(Click in this red rectangle to see SAS calculations for this example.)

This example two-factor problem can be defined as a 2x3 design because Factor A (Font) has two levels (Helvetica and Old English) and Factor B (Words/Line) has three levels (10, 20, and 30 words/line). It is a between-subjects design since 4 different subjects appeared in each of the 6 cells of the design.

Hypothetical data and ANOVA calculations are provided for this example problem on subsequent slides. Alternatively, Slater and Williges (2006) appendix provides the SAS procedure and results of this example problem using a statistical package.

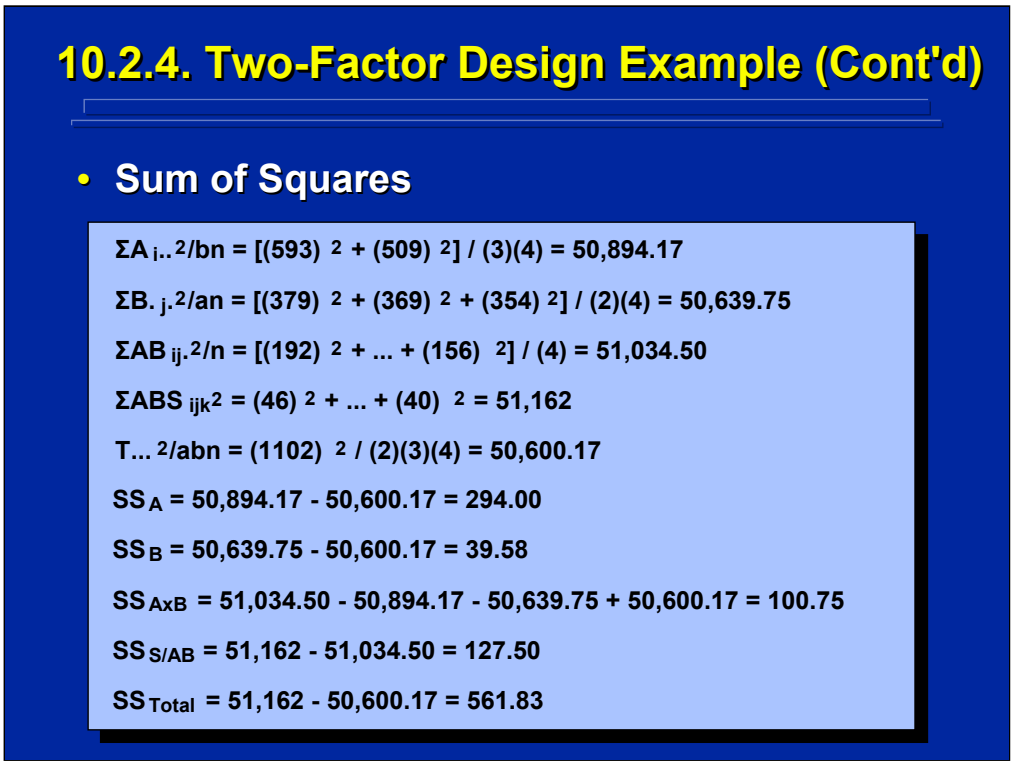
## 10.2.4. Two-Factor Design Example (Cont'd)

- **Data Matrix**

| Font (Factor A) | Words/Line (Factor B): 10 ($B_1$) | 20 ($B_2$) | 30 ($B_3$) | |
|---|---|---|---|---|
| Helvetica ($A_1$) | 46 | 49 | 50 | |
| | 50 | 52 | 47 | |
| | 49 | 54 | 49 | $A_{1..} = 593$ |
| | 47 | 48 | 52 | |
| | $AB_{11.} = 192$ | $AB_{12.} = 203$ | $AB_{13.} = 198$ | |
| Old English ($A_2$) | 47 | 39 | 35 | |
| | 46 | 44 | 42 | |
| | 50 | 38 | 39 | $A_{2..} = 509$ |
| | 44 | 45 | 40 | |
| | $AB_{21.} = 187$ | $AB_{22.} = 166$ | $AB_{23.} = 156$ | |
| | $B_{.1.} = 379$ | $B_{.2.} = 369$ | $B_{.3.} = 354$ | |
| | | | | $T... = 1102$ |

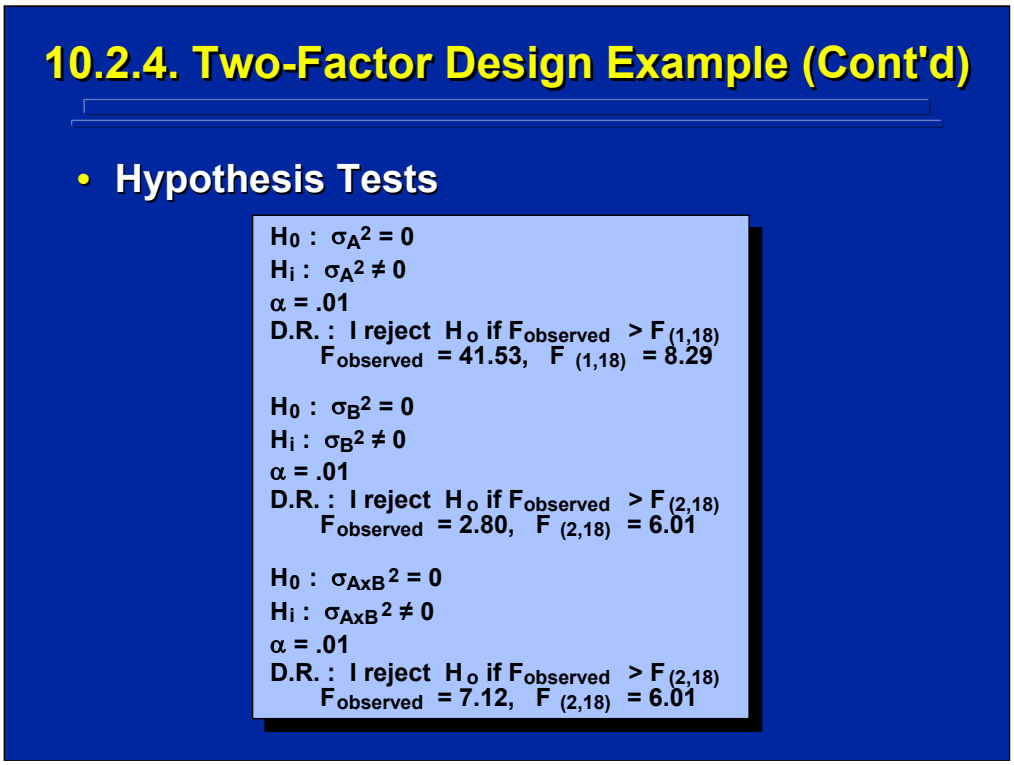
(Click in this red rectangle to see SAS calculations for this example.)

A hypothetical data matrix for a 2x3 between-subjects design is shown on this slide. Since the value of n is 4, the experimenter needs a total of 24 different subjects.



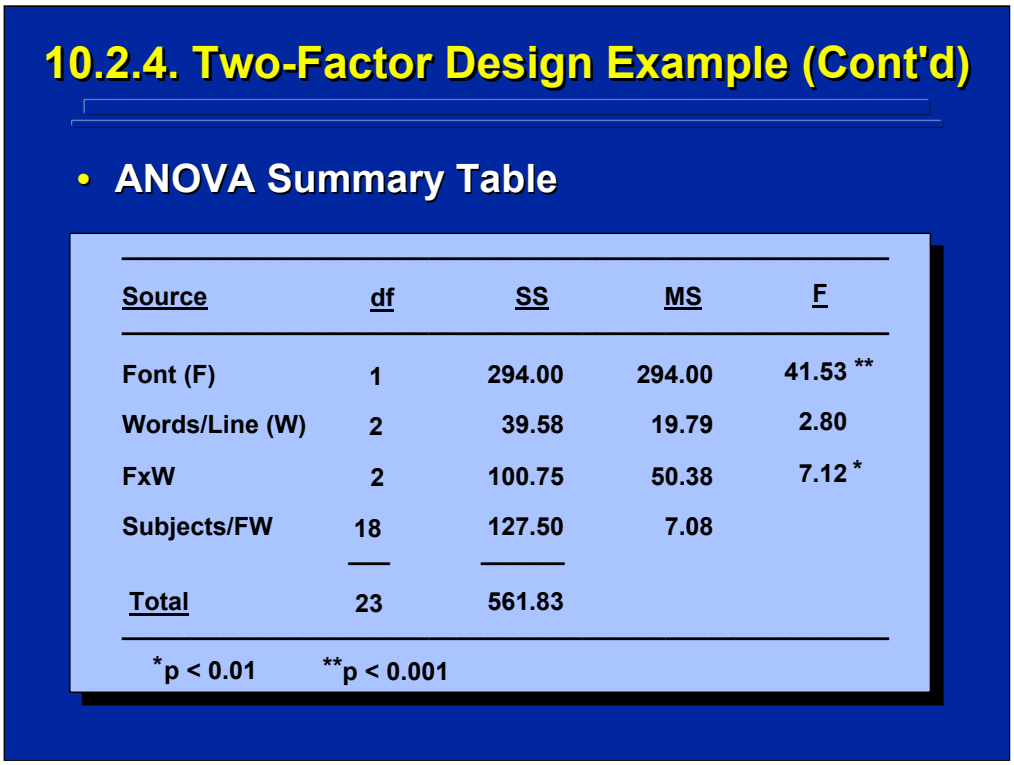

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the actual calculations of the sum of squares based on the SS computational formulae. The five components of the various formulae are listed on the top portion of this slide, and the SS of the various sources of the two-factor design are shown on the bottom portion of this slide.




(Click in this red rectangle to see SAS calculations for this example.)

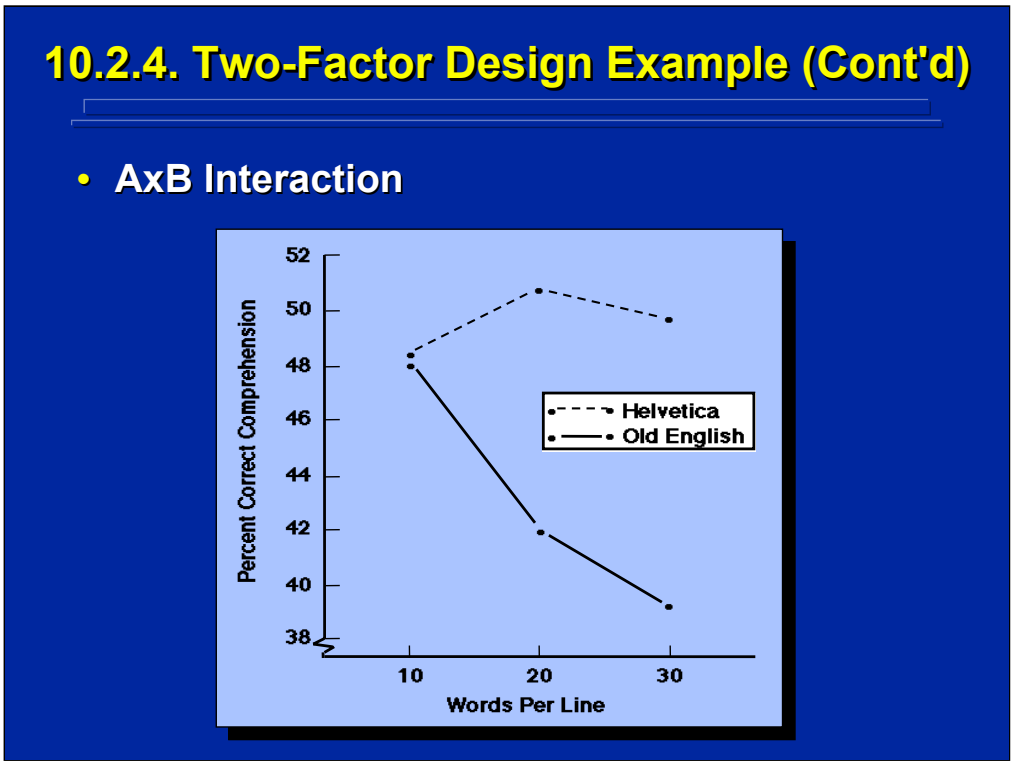
This slide shows the standard format testing the main effects of Factors A and B and the AxB interaction. Note that Factor A and the AxB interaction are significant at the 0.01 level. These tests form the basis of significance shown in the previous ANOVA Summary Table.




(Click in this red rectangle to see SAS calculations for this example.)

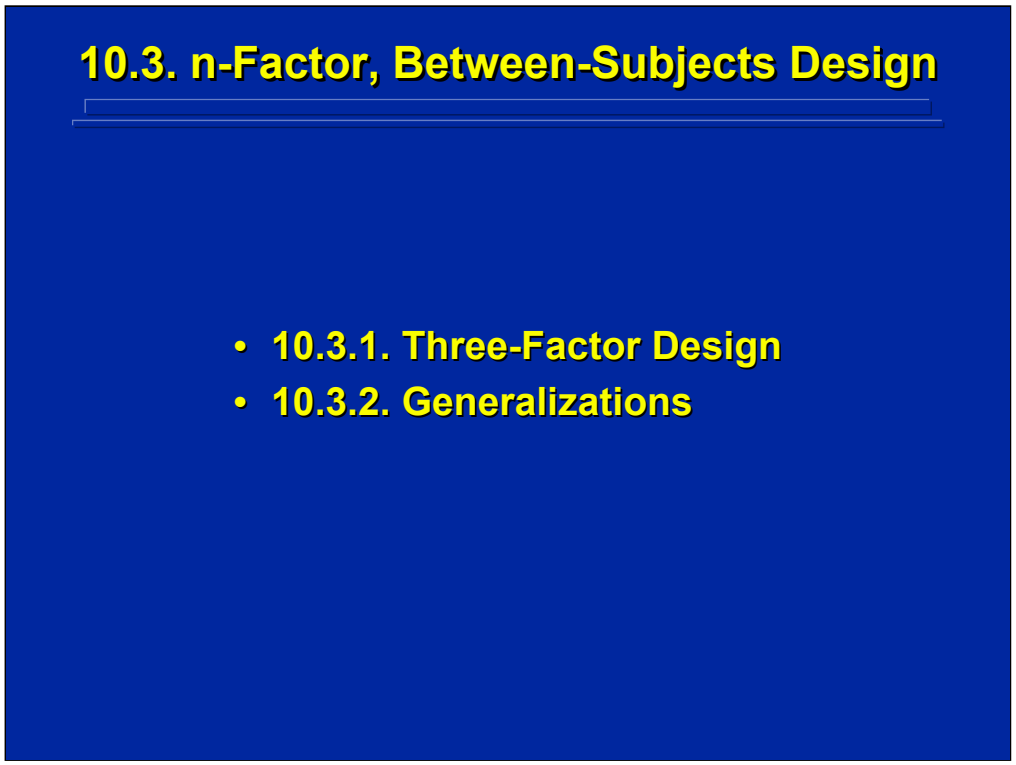
This slide shows the complete ANOVA Summary Table in terms of the actual factors investigated in the example problem. This is a more meaningful way of presenting the results to the reader rather than using A and B designations for the factors. The SAS results discussed by Slater and Williges (2006) for this problem provide the exact p-value of significance rather than just $p < 0.01$.

Note that Font is significant at the 0.001 level and the FxW interaction is significant at the 0.01 level when compared to the F tabled value (i.e. $F_{(1,18)} = 15.38$ and $F_{(2,18)} = 6.01$, respectively) as detailed in the hypothesis tests shown on the previous slide. Since there are only two fonts manipulated in the experiment, the researcher can conclude that overall the Helvetica font resulted in higher reading comprehension than the Old English font. The significant FxW interaction, however, needs further analysis in order to isolate the interaction effect. One could plot the interaction to determine the most likely simple effects that subsequently need to be supported by post hoc analyses on the interaction data. These analysis alternatives are discussed in Topic 11.




This figure shows the interaction of font type and words per line on reading comprehension. Even though on the average the Helvetica font results in significantly higher reading comprehension than the Old English font, the effect differs (interacts) depending upon the value of words/line ($p < 0.01$).

As shown in this figure, it appears that font type makes no difference in reading comprehension when only 10 words/line appear. The advantage of the Helvetica font in terms of reading comprehension seems to occur when 20 and 30 words/line are used. But, this apparent interaction effect needs to be verified analytically before making this interpretation. Various post hoc statistical analyses of interactions are discussed in Topic 10. The experimenter should always be careful to conduct post hoc statistical analyses to isolate the interaction effect rather than draw conclusions based solely on visual interpretations of the interaction graph.



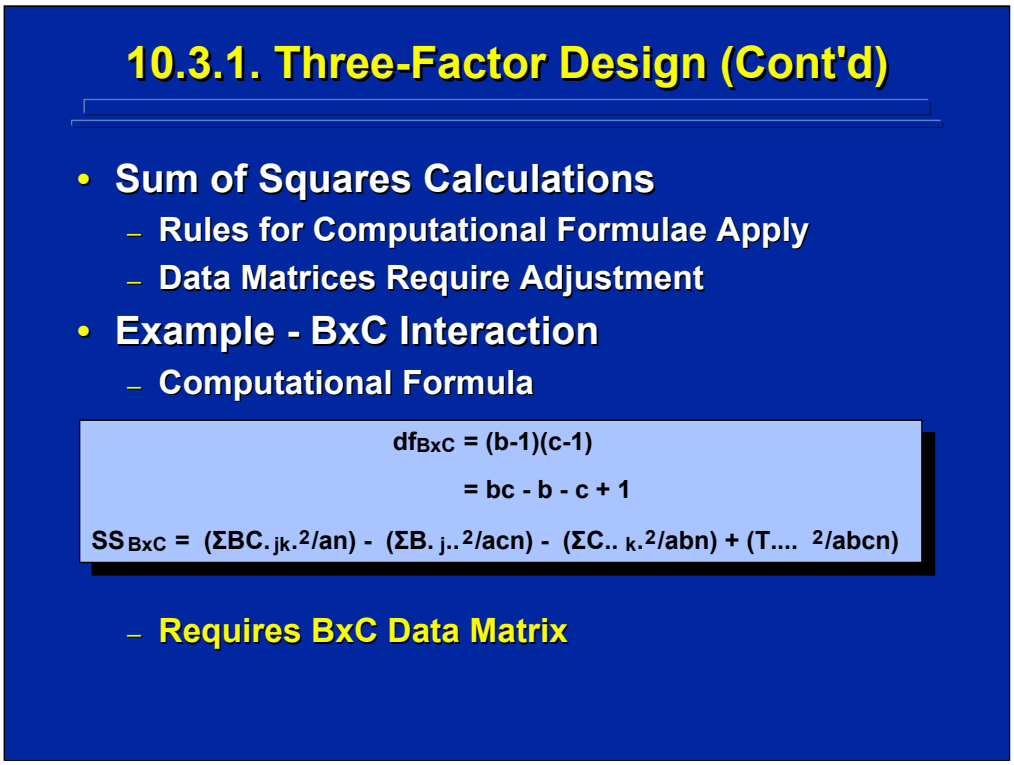

The procedures discussed for one-factor and two-factor between-subjects designs can be extended to higher-order factorial designs. This subsection first extends these procedures to a three-factor design and then generalizes them to n-factor between-subjects designs.

## 10.3.1. Three-Factor Design

$$Y_{ijklm} = \mu + \alpha_i + \beta_j + \delta_k + \gamma_{l(ijk)} + \alpha\beta_{ij} + \alpha\delta_{ik} + \beta\delta_{jk} + \alpha\beta\delta_{ijk} + \varepsilon_{m(ijkl)}$$

| Source | df | E(MS) | F |
|---|---|---|---|
| A | a-1 | $bcn\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$ | $MS_A/MS_{S/ABC}$ |
| B | b-1 | $acn\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$ | $MS_B/MS_{S/ABC}$ |
| C | c-1 | $abn\sigma_\delta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$ | $MS_C/MS_{S/ABC}$ |
| AxB | (a-1)(b-1) | $cn\sigma_{\alpha\beta}^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$ | $MS_{AxB}/MS_{S/ABC}$ |
| AxC | (a-1)(c-1) | $bn\sigma_{\alpha\delta}^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$ | $MS_{AxC}/MS_{S/ABC}$ |
| BxC | (b-1)(c-1) | $an\sigma_{\beta\delta}^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$ | $MS_{BxC}/MS_{S/ABC}$ |
| AxBxC | (a-1)(b-1)(c-1) | $n\sigma_{\alpha\beta\delta}^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$ | $MS_{AxBxC}/MS_{S/ABC}$ |
| S/ABC | abc(n-1) | $\sigma_\gamma^2 + \sigma_\varepsilon^2$ | |

This slide summarizes the statistical model, sources, degrees of freedom, E(MS), and F-ratios for a three-factor, between-subjects designs. All of these components were determined by the previously stated procedural rules and algorithms for ANOVA designs. Note that S/ABC is the error term for testing the three main effects, the three two-way interactions, and the single three-way interaction in this design.




The SS calculations for a three-factor design follow the same procedures as used in one- and two-factor designs. The same algorithm is used to generate the computational formulae.

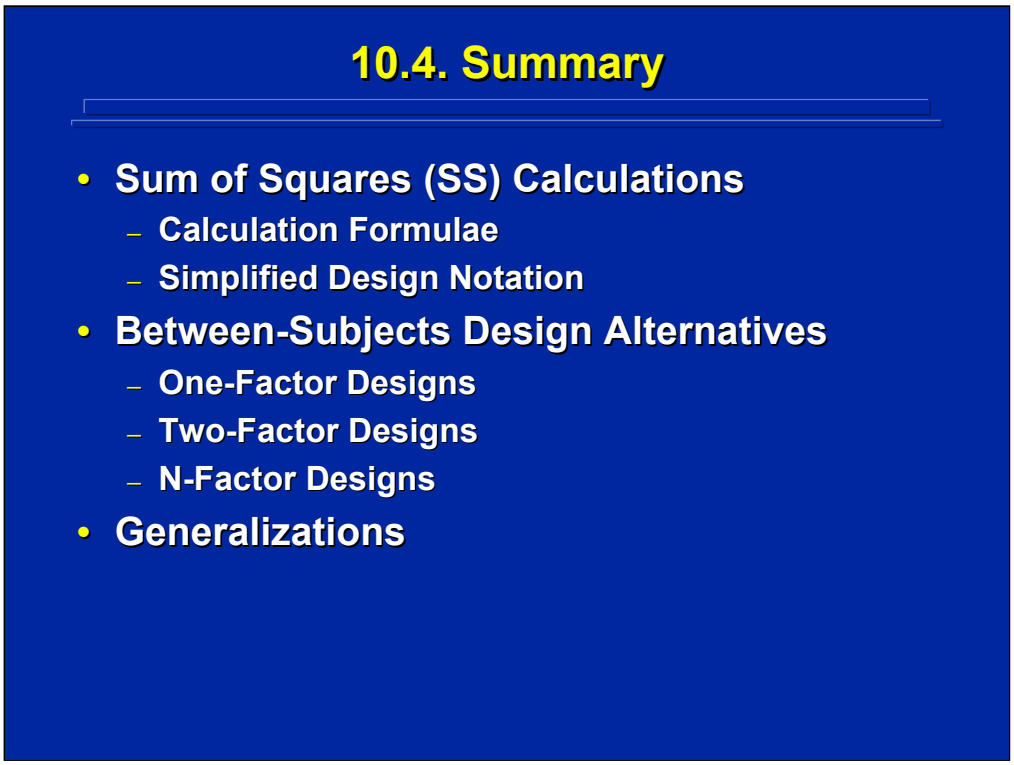
Obtaining the necessary totals to calculate interactions requires adjustment to the overall data matrix. For example, the SS formula for the BxC interaction is shown on this slide. Note that $BC_{.jk.}$ totals are needed in this formula. Consequently, the overall $ABCS_{IJKL}$ data matrix needs to be collapsed to a BxC interaction data matrix that sums across the levels of factor A and subjects in order to obtain the various $BC_{.jk.}$ totals.

## 10.3.2. Generalizations

- Can include any number of factors of interest.
- All rules, procedures, and algorithms apply.
- All factors of interest are crossed and can interact.
- Subjects are nested within all factors of interest.
- The subject effect is the error term for all F-tests.
  - Assumes subjects are random-effects.
  - Assumes factors of interest are fixed-effects.

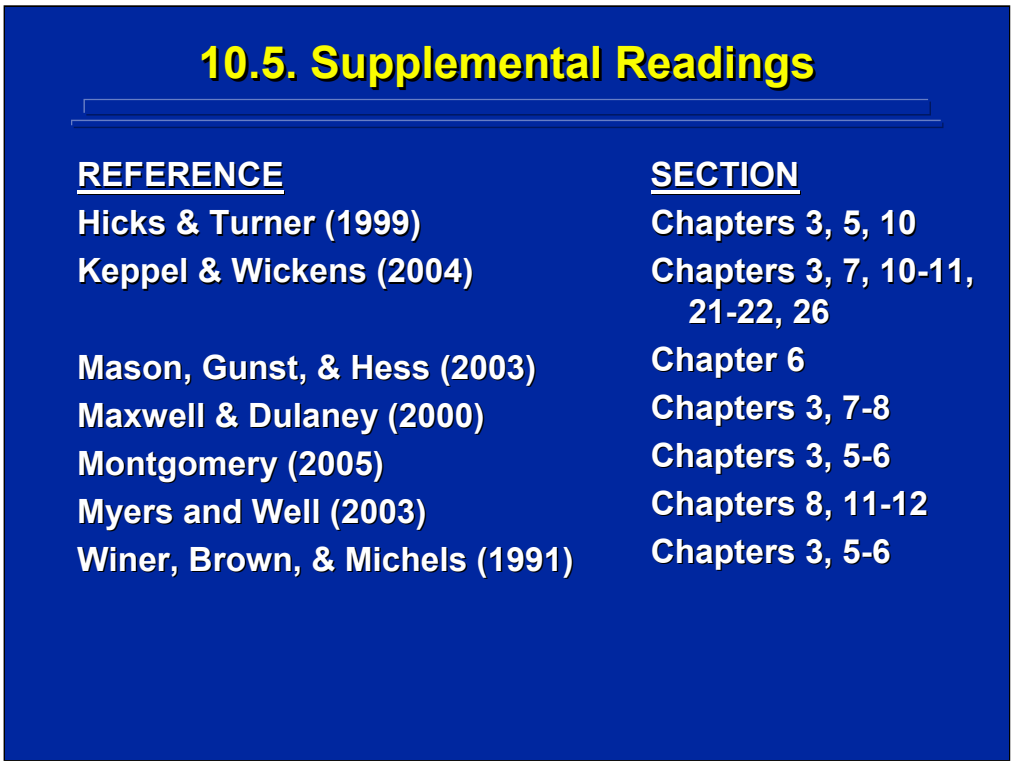
This slide provides the generalization of rules, procedures, and algorithms to any n-factor, between-subjects design. As the number of factors increases, more cells exist in the design requiring a larger number of different subjects to participate in the experiment. In addition, the number of interactions increases dramatically in factorial designs. For example, in a six-factor design there are six main effects and numerous 2-,3-,4-,5-,and 6-way interactions. In most human factors research, one is primarily interested in main effects and two-way interactions. Consequently, higher-order factorial designs are inefficient even though they can be easily constructed and analyzed.

Note that in any between-subjects design, the subject effect is the error term for all F-tests assuming the subjects effect is a random-effect variable and all the factors of interest are fixed-effects variables. If any of the factors of interest are truly random-effects variables, then the experimenter must specify the E(MS) to determine the appropriate error term for the resulting between-subjects design.




By way of summary, this topic covered three major concepts in ANOVA. First, a general algorithm is provided for constructing SS computational formulae based on a simplified design notation. These computational formulae can be generalized to any ANOVA design. Once the researcher understands the procedures for calculating SS, all the components are present for conducting the complete ANOVA.

Next, the major discussion in this topic is devoted to between-subjects ANOVA designs whether they be one-factor, two-factor, or n-factor designs. Computational examples are provided for both one-factor and two-factor between-subjects designs that can be easily extended to any n-factor design. Generalizations are provided that can be applied to any between-subjects design regardless of the number of factors included in the experiment.




The between-subjects design is the fundamental completely randomized ANOVA design and is covered in all experimental design textbooks addressing ANOVA. Appropriate chapters in common experimental design textbooks used by human factors researchers are listed on this slide. The chapters in Keppel and Wickens (2004) most closely follow the computational procedures covered in this topic.

# Topic 11. Analysis of Comparisons and Interactions

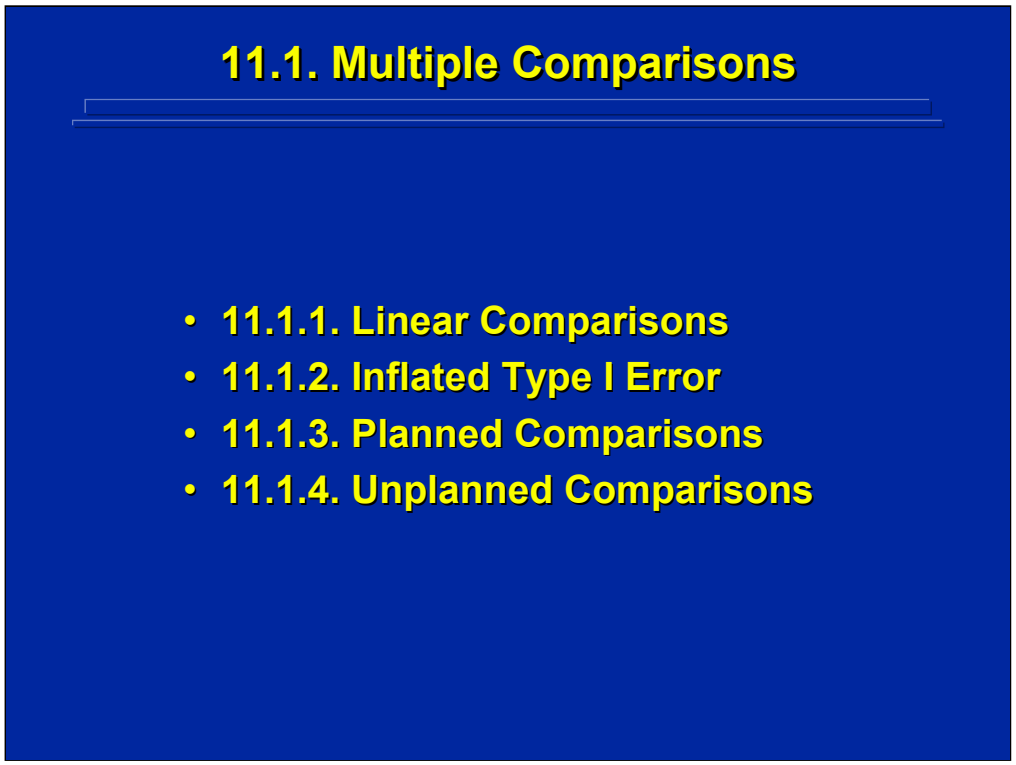

11.1. Multiple Comparisons
- 11.1.1. Linear Comparisons
- 11.1.2. Inflated Type I Error
- 11.1.3. Planned Comparisons
- 11.1.4. Unplanned Comparisons

11.2. Evaluating Interactions
- 11.2.1. Example Problem
- 11.2.2. Graphing Procedures
- 11.2.3. Simple Effects Test
- 11.2.4. Trend Analysis
- 11.2.5. Paired Comparisons
- 11.2.6. Interaction Evaluation Process

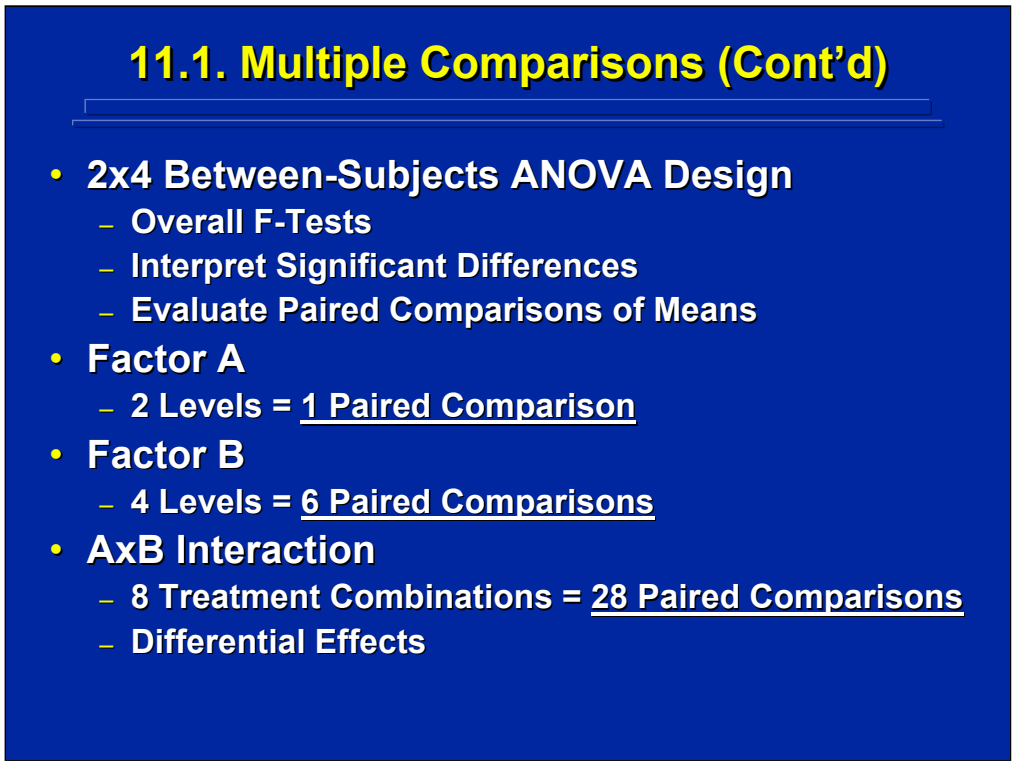
11.3. Summary

11.4. Supplemental Readings


This topic covers analytical techniques that can be used to isolate the form or nature of the main effects and interactions that are significant in the overall ANOVA. Basically these procedures deal with multiple paired comparisons of various treatment means. First, this topic covers comparisons and provides examples of both planned and unplanned comparison procedures. Second, special analytical procedures in addition to paired comparisons are covered, and example computations are provided for the analysis of interactions. Although these procedures are demonstrated using between-subjects designs, these same techniques are appropriate for within-subjects and mixed-factors designs. References to supplemental readings on comparisons are provided for additional readings in the major experimental design texts appropriate for human factors research.




Paired comparisons between treatment means are also referred to as contrasts in the statistical literature. These contrasts are linear comparisons that can be either planned a priori or unplanned comparisons that are conducted post hoc based on the results of the overall ANOVA. Since several paired comparisons are conducted on the same dataset, $\alpha$ error can inflate dramatically. Various analytical procedures have been developed to control for $\alpha$ error inflation in both planned and unplanned comparisons.




This slide shows the variety of paired comparisons that are present in a 2x4 factorial design. The ANOVA F-tests on the two main effects and the AxB interaction only show significant overall effects meaning that at least one of the paired comparisons is significantly different.

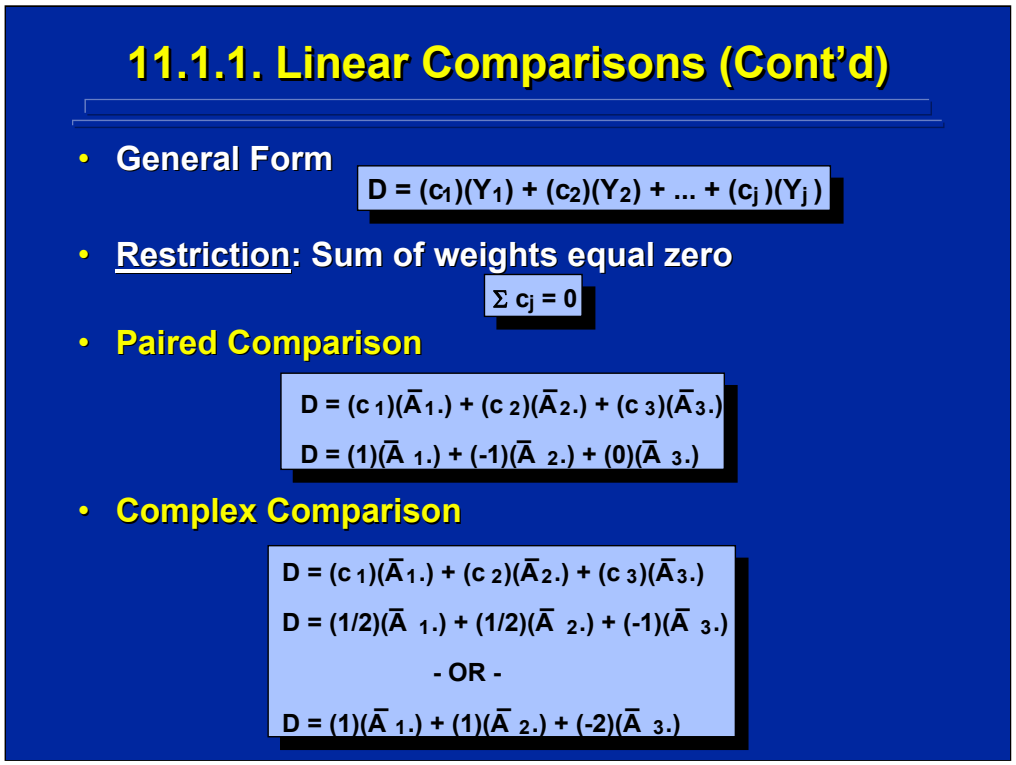
Factor A only has one paired comparison, because only two levels exist. Consequently, if the F-test for the main effect of Factor A is significant, no additional analysis is needed. The interpretation of a significant B main effect and the AxB interactions is not as simple since more than one paired comparison exists.

The total number of paired comparisons for any main effect or interaction can be determined by the combination counting rule. For example, the number of paired comparisons for Factor B equals the number of combinations of 4 means taken 2 at a time or 6 paired comparisons; whereas, the number of paired comparisons of the AxB interaction equals the number of combinations of 8 means taken 2 at a time or 28 paired comparisons. Since both the main effect of Factor B and the AxB interaction involve more than one paired comparison, additional analyses on the paired comparisons are needed to isolate the locus of the significant main effect and interaction.

## 11.1.1. Linear Comparisons

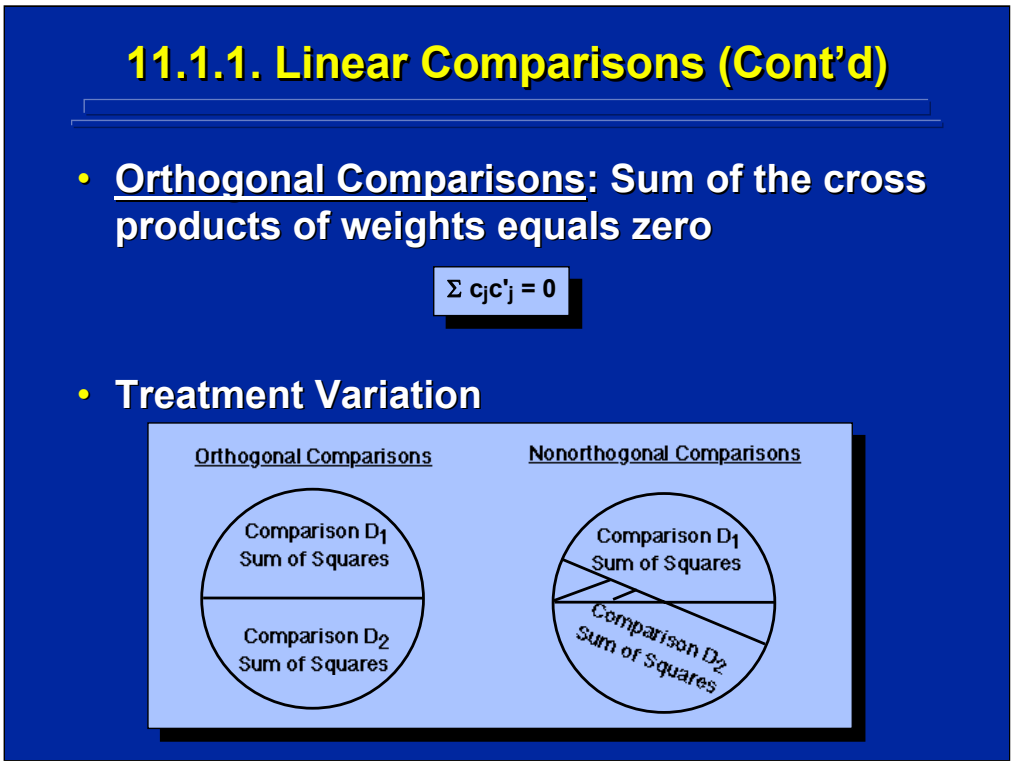
- Definition: A comparison is a difference between two means with the appropriate sign.
- Planned vs. Unplanned Comparisons
  - Planned Comparisons: A test for differences conducted instead of the overall F-test
  - Unplanned Comparisons: A post-hoc test to answer specific questions once an overall difference is determined by the F-test
- Paired vs. Complex Comparisons
  - Paired Comparison: A weighted combination of two means.
  - Complex Comparison: A weighted combination of several means.

Any linear comparison is defined as the algebraic difference between two means. These contrasts can be planned or unplanned prior to conducting the experiment and they can be a simple comparison of paired treatment means or complex comparisons of weighted combinations of several treatment means. In ANOVA, the experimenter is primarily interested in unplanned, paired comparisons to isolate significant main effects and interactions found in the overall ANOVA.

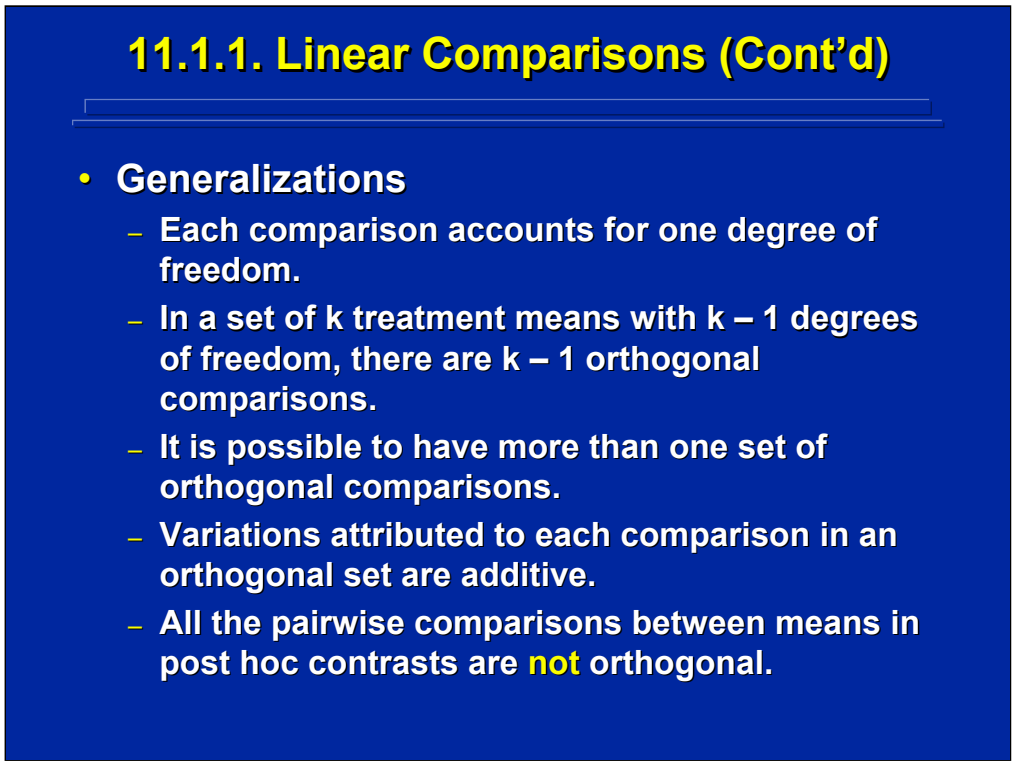



All linear comparisons or differences, D, of treatment means can be stated as weighted linear combinations with the restriction that the sum of the weights equals zero. This restriction is needed in order to keep any difference independent of the grand mean.

This slide shows an example of using a weighted combination to specify both paired and complex comparisons based on three treatment means in a one-way ANOVA where, for example, levels 1 and 2 may be experimental conditions and level 3 is the control condition. The paired comparison shown on the slide is the difference between levels 1 and 2 of Factor A since the weight of level 3 is 0. The complex comparison example shows the difference between the average of two treatment conditions, levels 1 and 2, and the mean of the control condition, level 3. Note that weights for complex comparisons are usually stated as integers as shown on the bottom of this slide to avoid rounding errors (i.e., 1, 1, and -2).

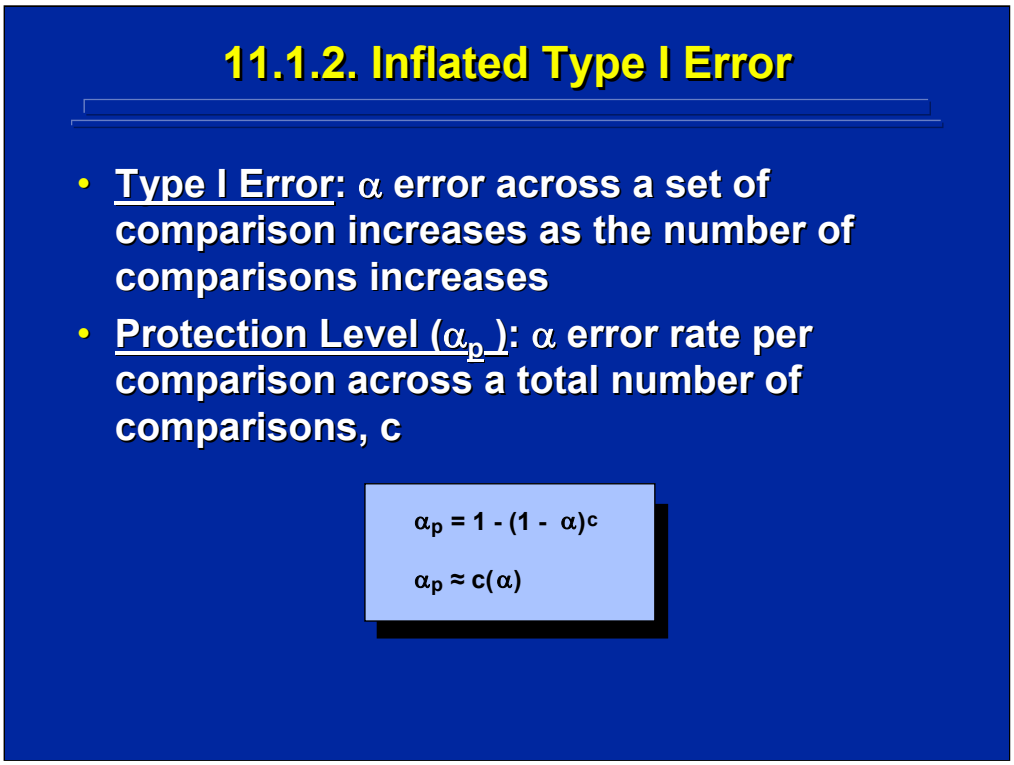



Two comparisons are defined as orthogonal if the sum of the cross products of the weights that comprise each difference, D, equals 0. If two comparisons are orthogonal, each comparison consists of independent sources of variation as shown in the Venn diagram. If the treatments are non-orthogonal, there is some overlap of the sum of squares as shown in the cross-hatched area of the Venn diagram on the right side of the slide. One uses both orthogonal and non-orthogonal comparisons in experimental design.

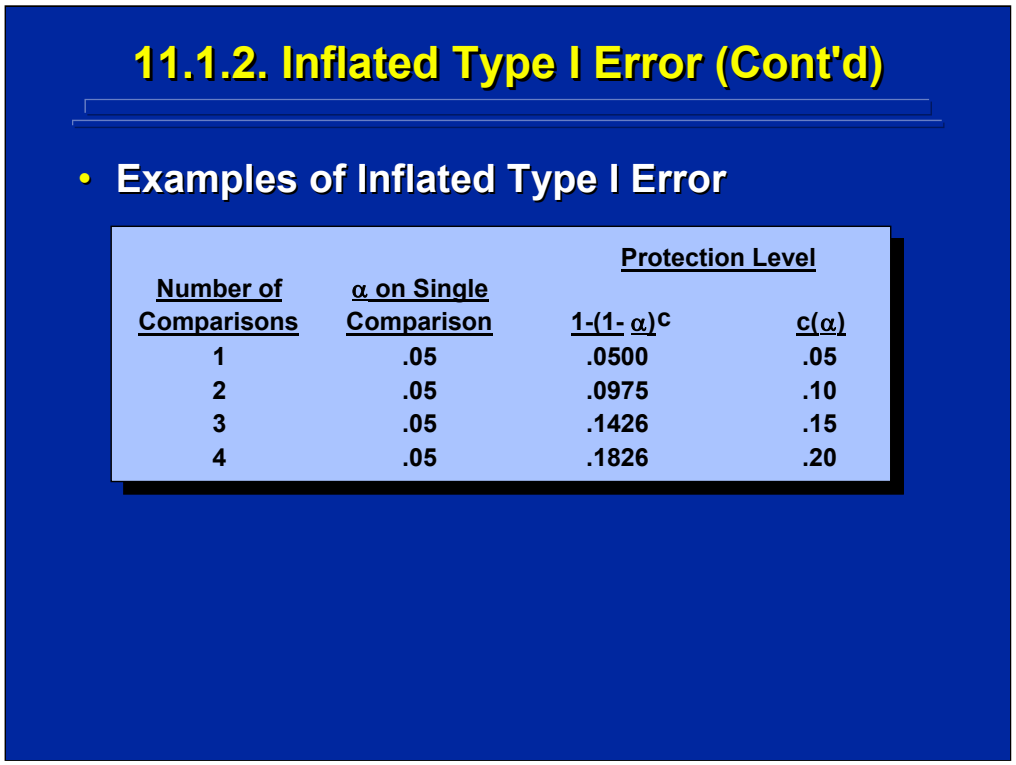



This slides summarizes characteristics of linear comparisons used in experimental design. Note that each comparison has 1 degree of freedom because a linear comparison is the difference between two means. Consequently there are k – 1 orthogonal comparisons among k treatment means, and several sets of orthogonal comparisons are possible. The sum of squares within an orthogonal set of comparisons are additive since the comparisons are independent contrasts.

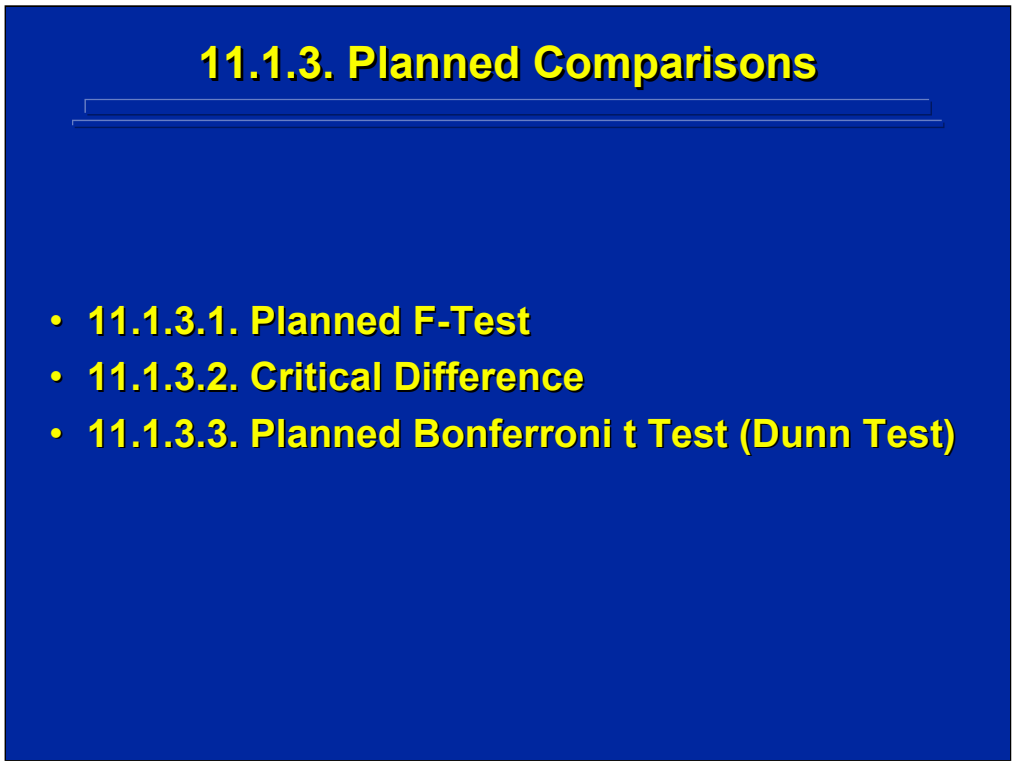
When all possible paired comparisons are used to isolate main effects and interactions of k treatments, these contrasts are not orthogonal because the number of comparisons in the set is greater than k – 1. Consequently, an experimenter primarily uses unplanned, non-orthogonal, paired comparisons to analyze main effects and interactions in ANOVA.



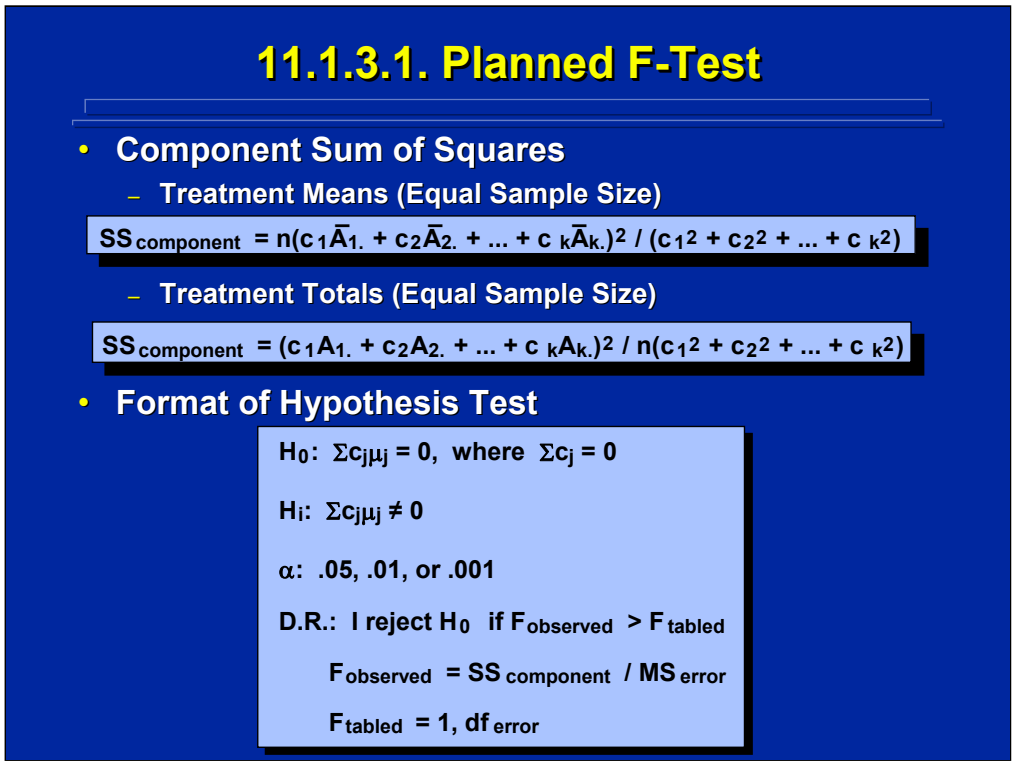

When several comparisons are made on the same set of data, the experimenter must be aware of inflated $\alpha$ error for each individual comparison. The protection level is the probability of at least one Type I error in a set of c independent comparisons. As this slide shows, the binominal formula specifies this inflated $\alpha$ error which can be approximated simply by $c(\alpha)$. Winer et al. (1991, pp. 153-158) and Maxwell and Delaney (2000, pp. 171-174) provide details on the probability of inflation of $\alpha$ error on sets of comparisons considered both experiment wise (i.e., comparisons across the entire experiment) and family wise (i.e., comparisons to isolate main effects and interactions) in ANOVA.




This slide shows various examples of inflated Type I error of hypotheses tested at the 0.05 level of significance. When 4 independent comparisons are conducted on the same data set, the probability of finding at least one significant difference by chance inflates from the original 0.05 to 0.20. Note that the c($\alpha$) approximation is close to the binomial solution.




Although human factors researchers are primarily interested in unplanned comparisons to evaluate main effects and interactions in ANOVA, there are occasions when a set of comparisons is planned a priori. This subtopic demonstrates component SS calculations for conducting planned comparisons, a general form for conducting these comparisons expressed as critical differences, and the Bonferroni t test as one technique for controlling inflated $\alpha$ error across a set of planned comparisons.




The top portion of this slide shows the general SS formulae for planned, weighted linear comparisons of treatments with equal sample size, n. Both the SS component formulae based on treatment means and totals are provided. They are equivalent.

The general format for conducting a statistical hypothesis test on a planned comparison is shown on the bottom portion of this slide. Note that an F-test is used based on the $MS_{error}$ of the main effect or interaction relevant to the planned comparison. A comparison has only 1 degree of freedom, hence the $F_{Observed}$ value is equivalent to just $SS_{Component}$ divided by $MS_{Error}$.

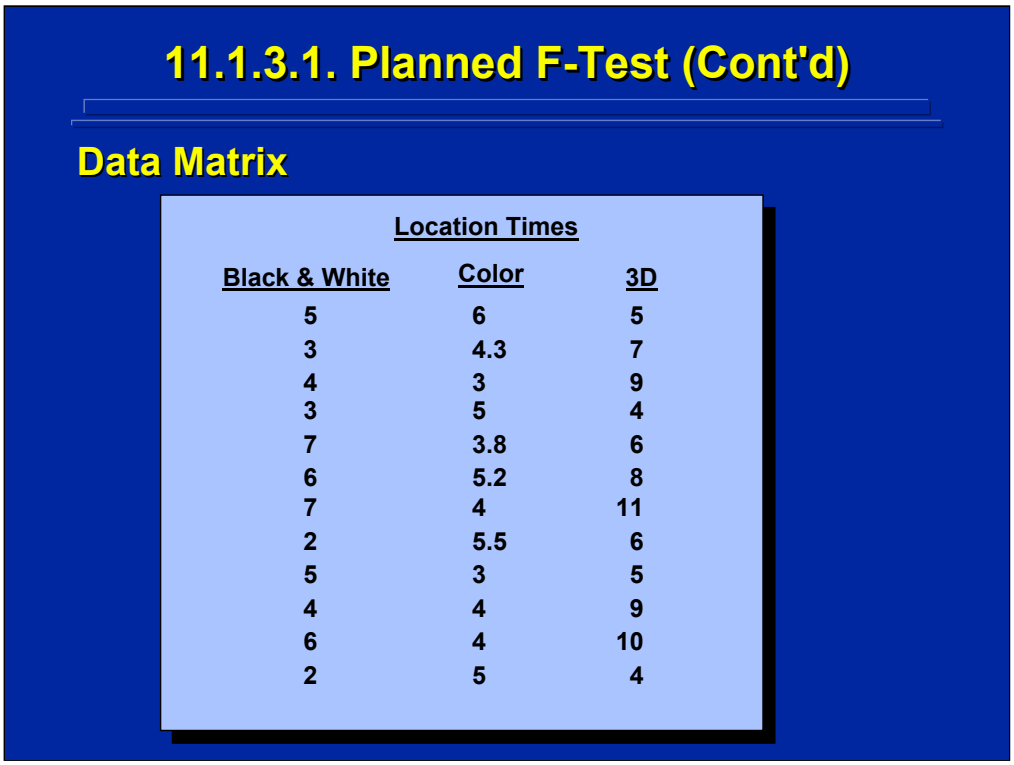
## 11.1.3.1. Planned F-Test (Cont'd)

- **Example Problem: The average number of seconds for 12 soldiers to locate a position on a standard black and white navigational map was compared to 12 other soldiers using an experimental colored map, and 12 other soldiers using an experimental 3-D map. Four tests of significant differences in location time were planned: standard versus color, standard versus 3-D, color versus 3-D, and standard versus the average of color and 3-D maps. Which differences were significant ($p < 0.05$)?**

(Click in this red rectangle to see SAS calculations for this example.)

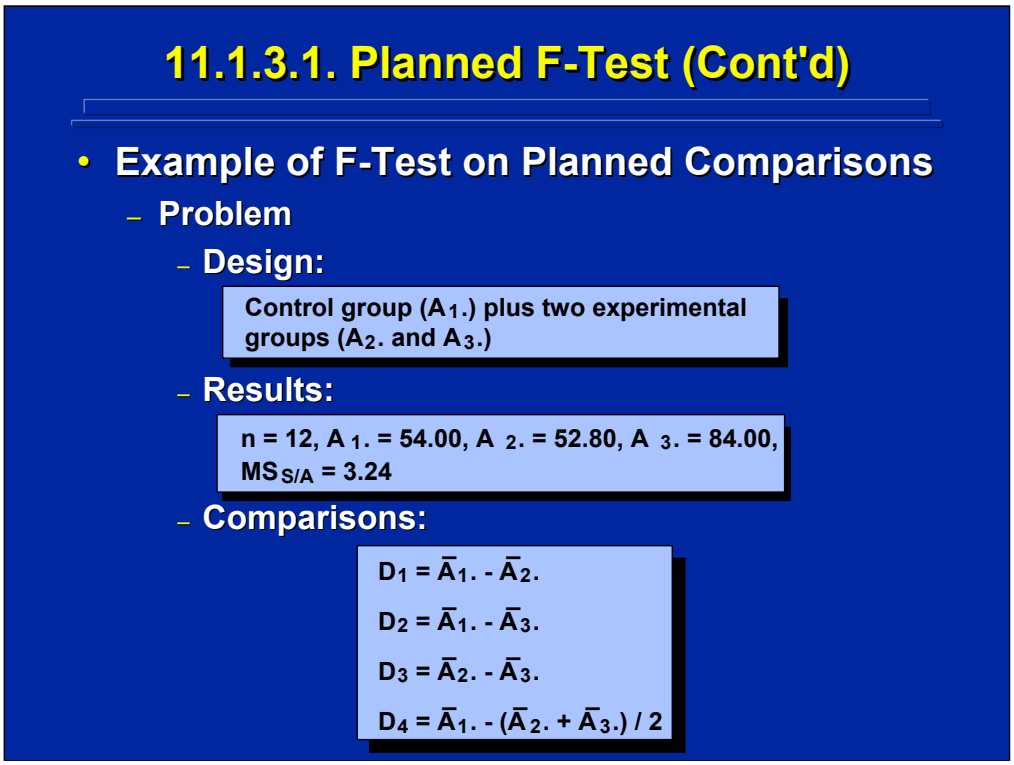
This slide describes four planned comparisons based on a one-way, between-subjects experimental design that has three treatment levels. Notice that the first three comparisons are simple, paired planned comparisons; whereas the fourth comparison is a complex, planned comparison. Since these comparisons were planned a priori, an overall F-test is not necessary. The one-way ANOVA, however, provides the $MS_{Error}$ value needed in each of the four contrasts tested at the 0.05 level of significance.

This reference material demonstrates the hand calculations for these four comparisons. The Slater and Williges (2006) appendix provide the SAS results for these contrasts using a statistical package.



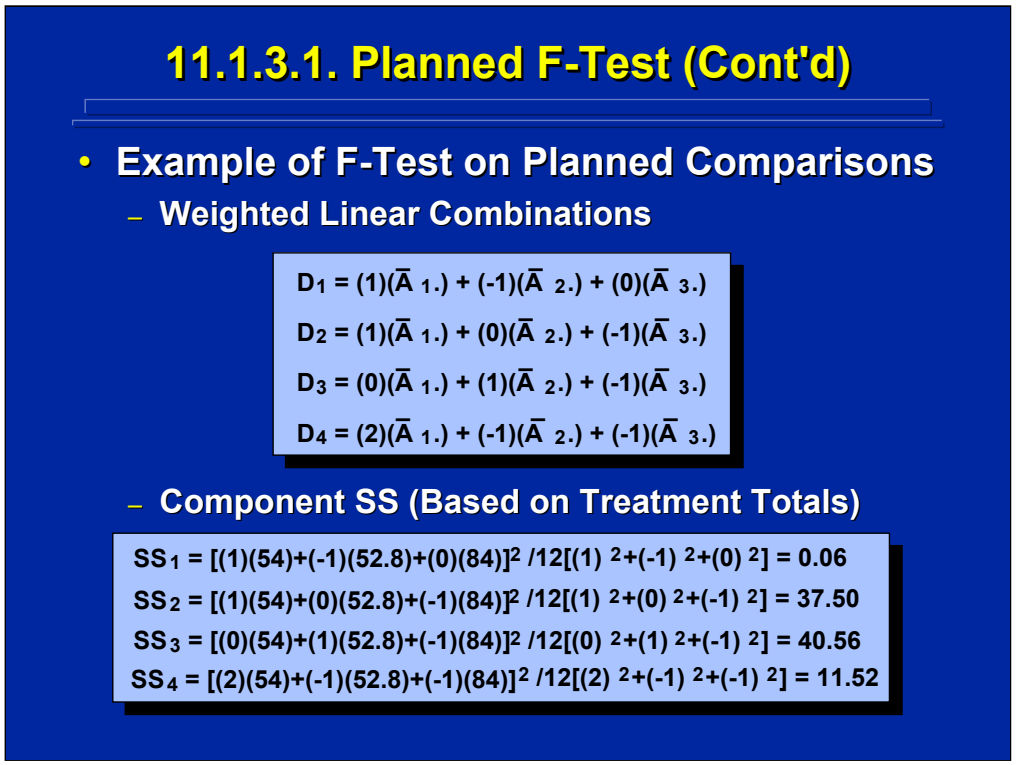

(Click in this red rectangle to see SAS calculations for this example.)

This slide summarizes a hypothetical data set of location times in seconds for the three map display comparisons in the example problem described on the previous slide. Each map was used by 12 different soldiers resulting in a total of 36 soldiers who participated in the experiment.




(Click in this red rectangle to see SAS calculations for this example.)

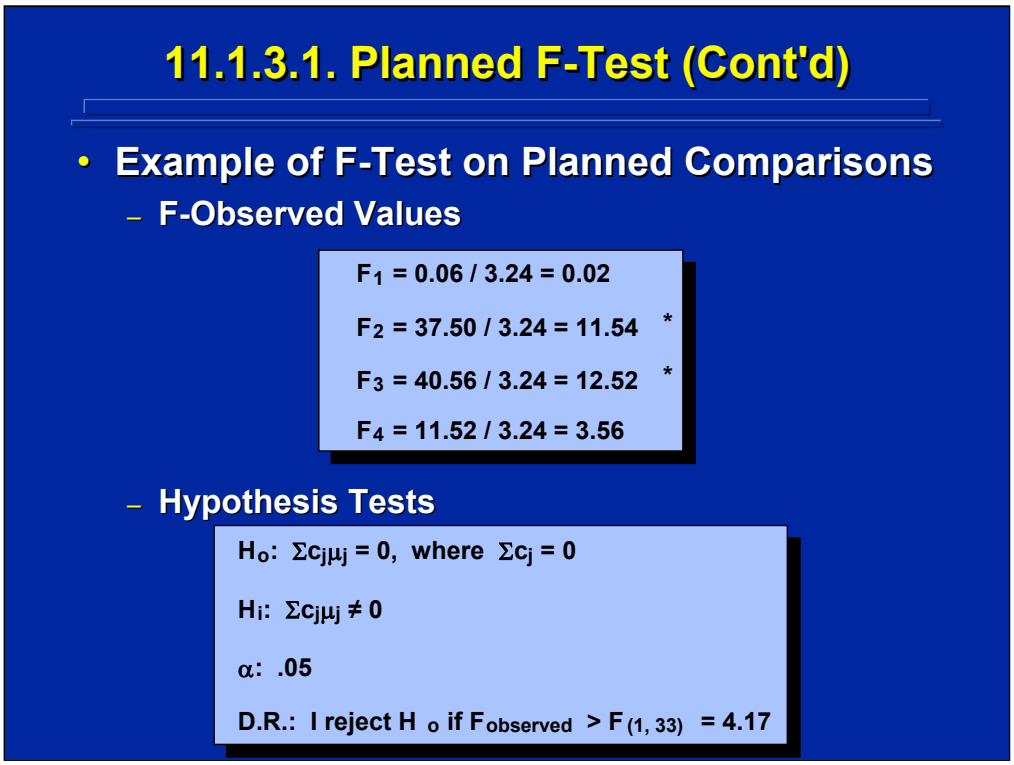
This example problem can be considered as a simple one-way ANOVA between-subjects design in which the standard map is the control condition and both the color and 3-D maps are the experimental conditions. Hypothetical results are shown in the middle portion of this slide in terms of treatment totals. The four planned comparisons are stated at the bottom of the slide. The first three are simple, paired comparisons, and the fourth is a complex comparison.




(Click in this red rectangle to see SAS calculations for this example.)

The weighted linear comparisons using integer weights are listed for each of the four planned contrasts in the top portion of this slide. Note that the sum of the weights equal 0 for each of the four comparisons. As a group, these comparisons are not orthogonal because the four degrees of freedom of the contrasts are more than the two degrees of freedom of the main effect of the three treatments. By calculating the sum of the cross products of pairs of contrasts, one can determine that $D_3$ and $D_4$ are orthogonal contrasts. The other pairs of planned comparisons consider overlapping sources of information.

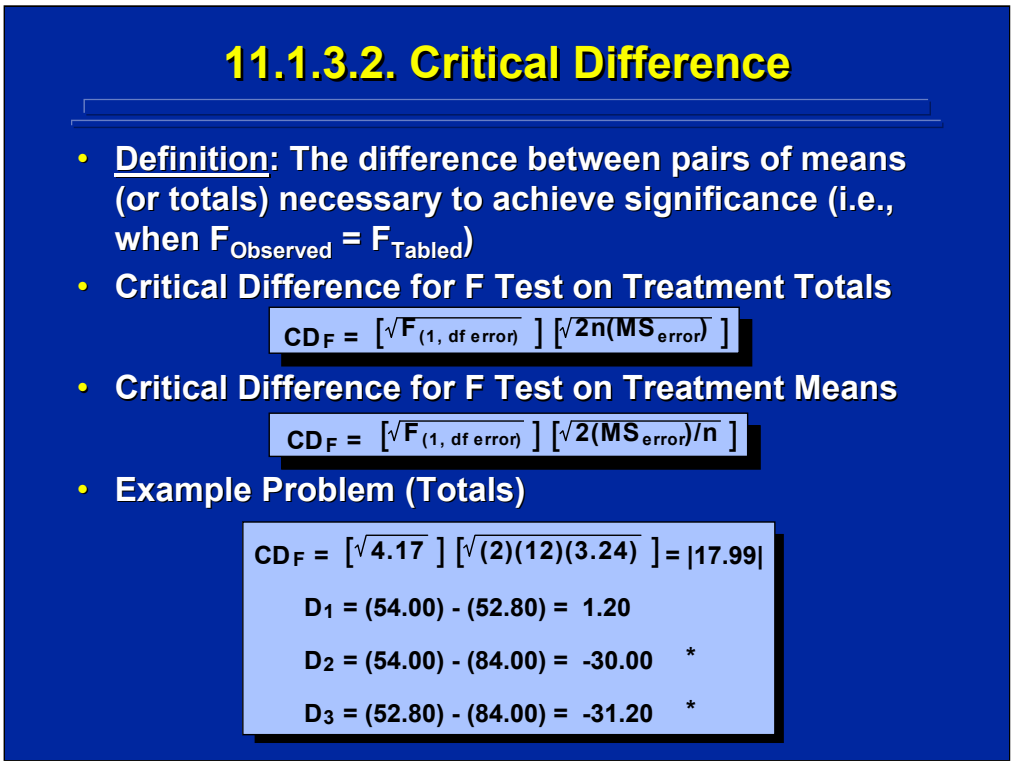
The bottom portion of the slide shows the component SS calculations based on treatment totals for each of the four planned comparisons. These SS values are used in the subsequent F-test on each planned comparison.




(Click in this red rectangle to see SAS calculations for this example.)

The F-ratios for each of the four planned contrasts are shown on the top portion of this slide, and the standard format for testing each hypothesis is shown the bottom portion. Since the $F_{Tabled}$ value is 4.17 for each comparison, only $D_2$ and $D_3$ are significant at the 0.05 level. Consequently, the difference between the black and white navigational display mean and the 3-D navigational display mean is significant (i.e., $D_2$), and the difference between the color display mean and 3-D display mean (i.e., $D_3$) is significant.

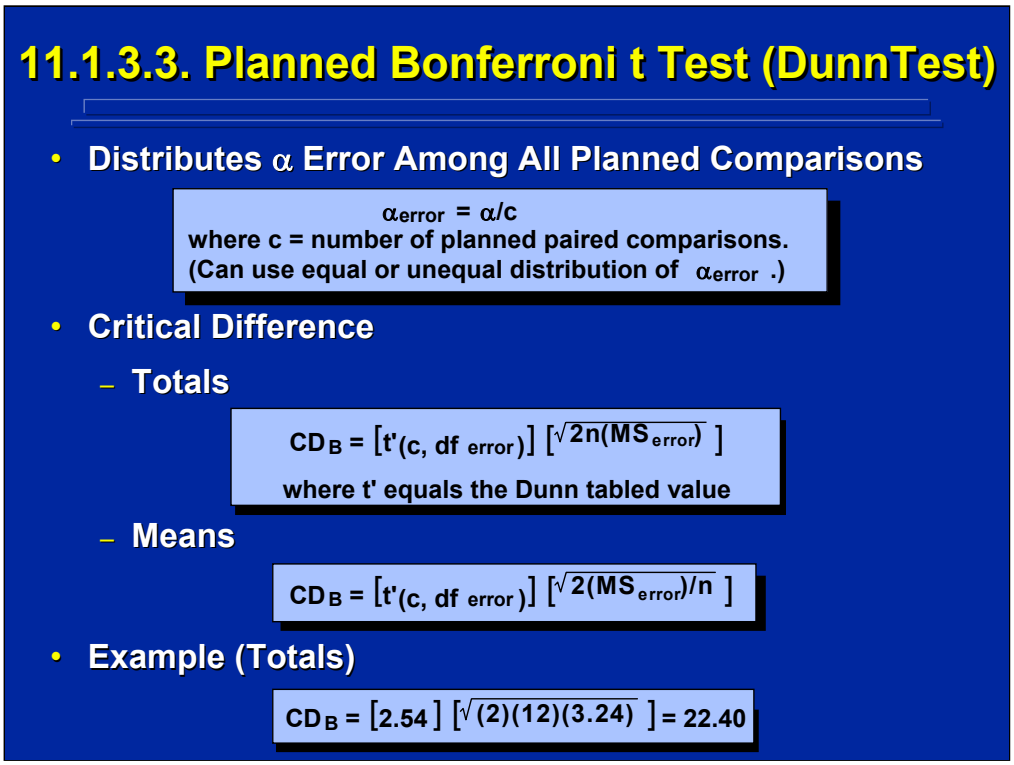
Remember that the F-test with one degree of freedom is identical to the t test. Slater and Williges (2006) show the results of t-tests on these contrasts using SAS.




(Click in this red rectangle to see SAS calculations for this example.)

An alternate way of calculating paired comparisons is to state the critical difference in terms of means or totals between two treatments that is necessary to achieve statistical significance. The critical difference formula, $CD_F$, for the F-tests on comparisons is shown on this slide for both treatment means and totals assuming equal sample size. For the paired comparisons in the example problem, the critical difference for treatment totals is 17.99 as shown on the slide.

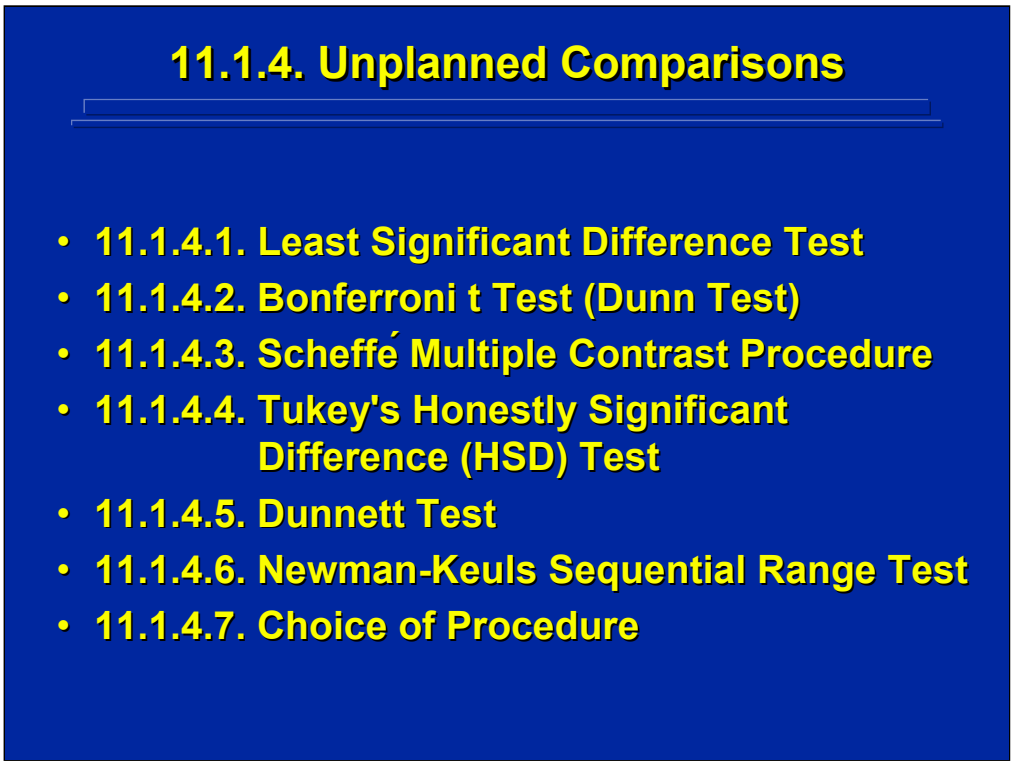
To test for significance, the experimenter only needs to calculate the difference between a pair of treatment means or totals and determine if the absolute value is equal to or greater than the critical difference. If so, the paired comparison is significant. Note that both $D_2$ and $D_3$ are significant as determined by the component SS calculation. The critical difference calculation is much easier, and this approach is used in discussing all subsequent paired comparison alternatives. To distinguish among these various paired comparison alternatives, each critical difference (CD) is given a unique subscript to designate that analytical alternative. For example, $CD_F$ shown on this slide designates a comparison based on the F-statistic sampling distribution.




(Click in this red rectangle to see SAS calculations for this example.)

The F-test used on the planned comparisons for this example did not control for inflated $\alpha$ error. One way to provide a control is to use a Bonferroni t test or a Dunn test. It takes the overall $\alpha$ error and divides it by the number of comparisons. The Bonferroni t test uses the table value, t' as presented in Appendix Table D.16 in Winer et al. (1991). The critical difference formula based on either means or the totals are shown in the middle portion of this slide.

Since the experimenter controls for inflated Type I error across all comparisons when using the Bonferroni t, the difference between a pair of treatments must be larger than with the F-test that uses an overall $\alpha$ value. For this example problem the $CD_B$ shown on this slide is larger than the $CD_F$ shown on the previous slide (i.e., 22.40 versus 17.99).




For unplanned, multiple comparisons several different alternatives are available to control for inflated $\alpha$ error. Many of these alternatives are discussed in detail by Winer, et al. (1991) in Chapter 3, pp.153-197, and Maxwell and Delaney (2000) in Chapters 5 and 6. The basis of control and the critical difference formulae of six alternative paired comparison procedures often used in human factors research are summarized in this subsection along with a discussion of the appropriate choice of procedure to use for a particular experiment.

## 11.1.4. Unplanned Comparisons (Cont'd)

- **Use**
  - **Post Hoc Analyses for Additional Data Investigation**
  - **Isolate Significant Main Effects and Interactions of Overall ANOVA**
- **Approach**
  - **Critical Difference Formulae for Means and Totals**
  - **Uses Error Term from Overall Analysis**
  - **Common Example: Single Factor, Between-Subjects Design**

Unplanned comparisons are used primarily in human factors research as a means of investigating significant effects found in the overall ANOVA to isolate the locus of main effects and interactions. Since these involve post hoc analyses, all these comparisons are unplanned.

The critical difference formulae for both means and totals are provided for each of the alternative unplanned comparison procedures. An $MS_{Error}$ term appears in each critical difference formulae. The appropriate overall ANOVA error term for the main effect or interaction being evaluated is used as the $MS_{Error}$ in the critical difference formula. To facilitate comparisons among alternative procedures and to provide a computational example, each unplanned comparison procedure uses the results of a significant main effect of a one-way, between-subjects ANOVA design.

## 11.1.4. Unplanned Comparisons (Cont'd)

- **Example Problem: Proprioceptive, visual, sound, and voice modes of presenting information were evaluated by 24 soldiers. One of these four modes of information was randomly assigned to 6 soldiers using wearable computers during training maneuvers. There was an overall significant mode difference in minutes to complete the training maneuver ($p < 0.05$). Which communication modes were significantly different from each other?**

(Click in this red rectangle to see SAS calculations for this example.)

This slide describes a one-factor, between-subjects ANOVA design that is used as the example problem for each unplanned comparison procedure described in this reference material. Note that this design has four levels. The SAS procedures for conducting each of the unplanned comparison procedures are described in the Slater and Williges (2006) appendix.

## 11.1.4. Unplanned Comparisons (Cont'd)

- **Data Matrix**

| | Mode of Presentation (A) | | | |
|---|---|---|---|---|
| | Proprioceptive ($A_1$) | Visual ($A_2$) | Voice ($A_3$) | Sound ($A_4$) |
| | 10 | 10 | 15 | 13 |
| | 13 | 10 | 15 | 14 |
| | 14 | 13 | 16 | 12 |
| | 14 | 14 | 17 | 18 |
| | 15 | 12 | 16 | 16 |
| | 17 | 15 | 18 | 19 |
| Total | 83 | 74 | 97 | 92 |
| Mean | 13.83 | 12.33 | 16.17 | 15.33 |

(Click in this red rectangle to see SAS calculations for this example.)

This slide provides a hypothetical data set for the example problem described on the previous page. This data set is used to illustrate each unplanned comparison procedure. Since this is a between-subjects design, six different soldiers (i.e., n = 6) were observed in each of the four modes of presentation to yield a total of 24 soldiers used in the experiment.

## 11.1.4. Unplanned Comparisons (Cont'd)

- **Common Example: Single Factor, Between-Subjects Design**

**ANOVA Summary Table**

| Source | df | SS | MS | F |
|---|---|---|---|---|
| A | 3 | 51.50 | 17.17 | 3.64* |
| S/A | 20 | 94.33 | 4.72 | |
| Total | 23 | 145.83 | | |

*($p < 0.05$) where, $F_{(3,20)} = 3.10$

(Click in this red rectangle to see SAS calculations for this example.)

This slide presents the ANOVA Summary Table for the overall analysis conducted on the example problem data. There are four levels of Communication Mode (Factor A), and n is equal to six, which gives a total of 23 degrees of freedom across the 24 observations. Since Factor A is significant, at least one of the paired differences among the four levels of Factor A must be significant. Subsequent paired comparisons are needed to determine specifically which differences are significant. These resulting paired comparisons are non-orthogonal, and they are unplanned since they are conducted only after establishing the overall significant main effect.

## 11.1.4. Unplanned Comparisons (Cont'd)

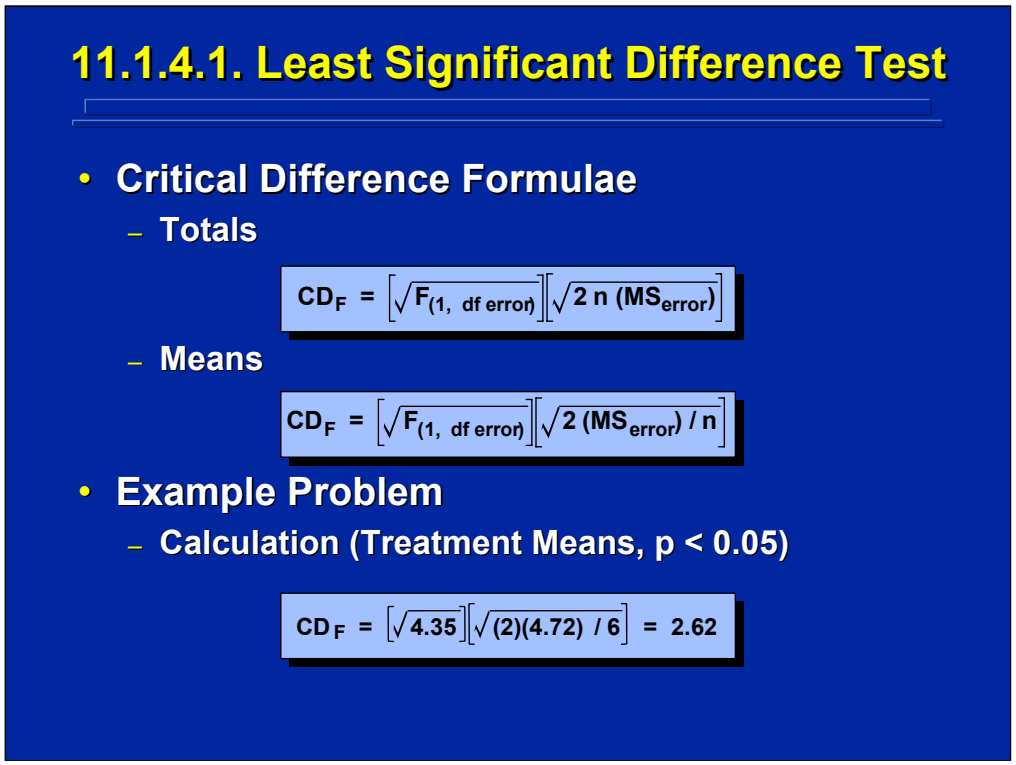
- **Common Example: Single Factor, Between-Subjects Design**

**Differences of Ordered Treatments Means**

| | | Increasing Order | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| Treatments, $(A_i)$ | | $A_2$ | $A_1$ | $A_4$ | $A_3$ |
| Means, $(\bar{A}_i)$ | | 12.33 | 13.83 | 15.33 | 16.17 |
| | $a_2$ | ---- | 1.50 | 3.00 | 3.83 |
| Increasing Order | $a_1$ | | ---- | 1.50 | 2.33 |
| | $a_4$ | | | ---- | 0.83 |
| | $a_3$ | | | | ---- |

(Click in this red rectangle to see SAS calculations for this example.)

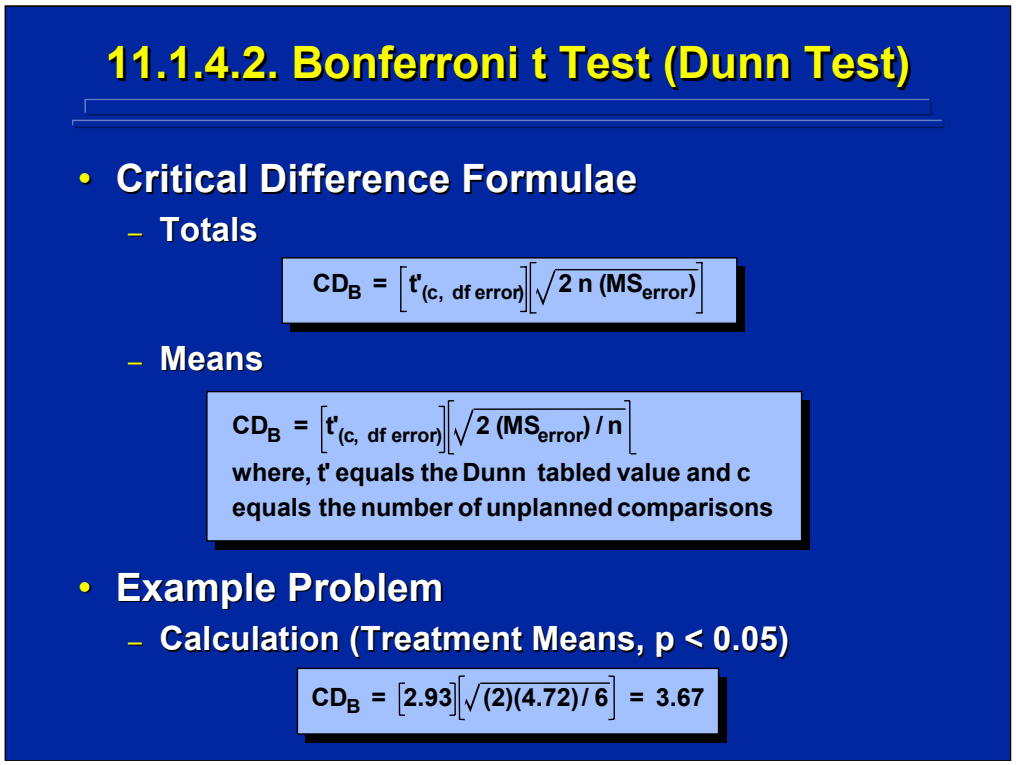
This slide depicts a table of ordered differences between pairs of the four treatment means in the example problem going from smallest to largest mean. Six paired differences are possible from the combination of four treatments. Any difference in paired means shown in this table that is larger than the critical difference calculated from the analytical procedure chosen is significant. Based on the overall F test, the experimenter knows the largest difference, 3.83 (or the difference between $a_2$ and $a_3$), is significant but any other significant differences must be determined through subsequent unplanned comparisons.




(Click in this red rectangle to see SAS calculations for this example.)

The Least Significant Difference (LSD) test is really a t-test that makes no correction for $\alpha$ error. This is the least stringent of the paired comparison procedures and, consequently, results in the largest number of significant differences.

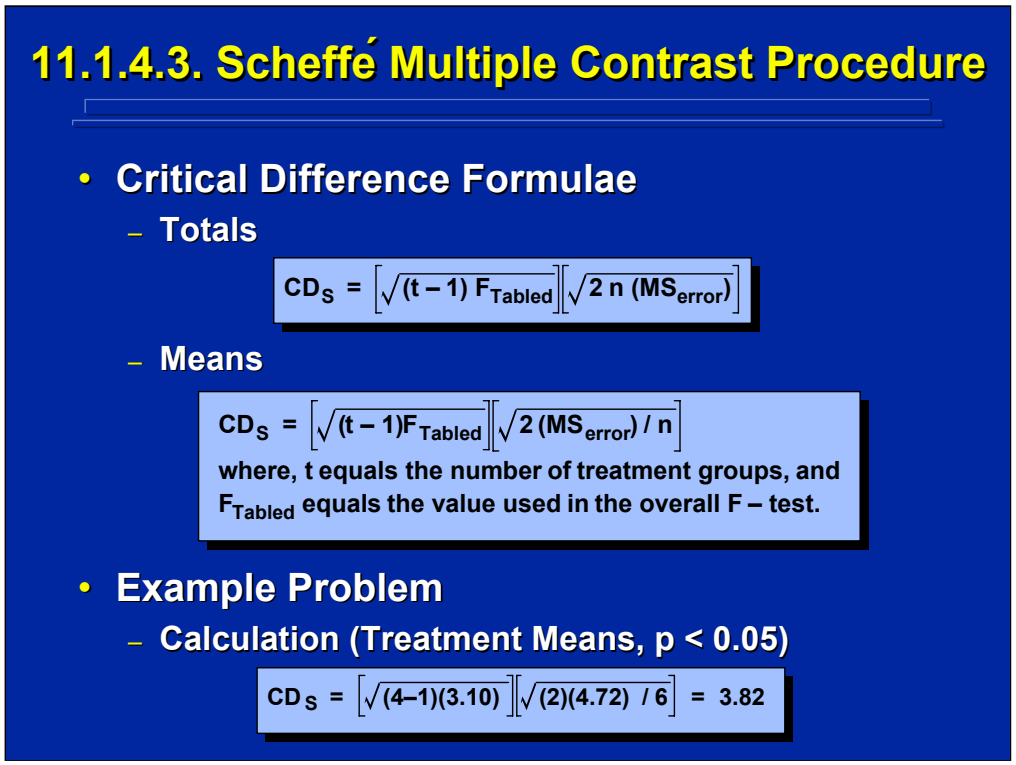
Using the critical difference formula for means presented on this slide for the LSD test, a difference of 2.62 between any pair of means is significant. Consequently, the differences between means $a_2$ and $a_3$ and means $a_2$ and $a_4$ are significant at the 0.05 level.




(Click in this red rectangle to see SAS calculations for this example.)

The Bonferroni t Test is the same test that was discussed for planned comparisons. Recall that $\alpha$ error is distributed across all comparisons, c. In this example c = 6, the number of unplanned, post hoc paired comparisons. The tabled t' value (i.e. $t'_{(6,20)} = 2.93$) is from the Bonferroni table presented in Appendix Table D.15 in Winer et al. (1991).

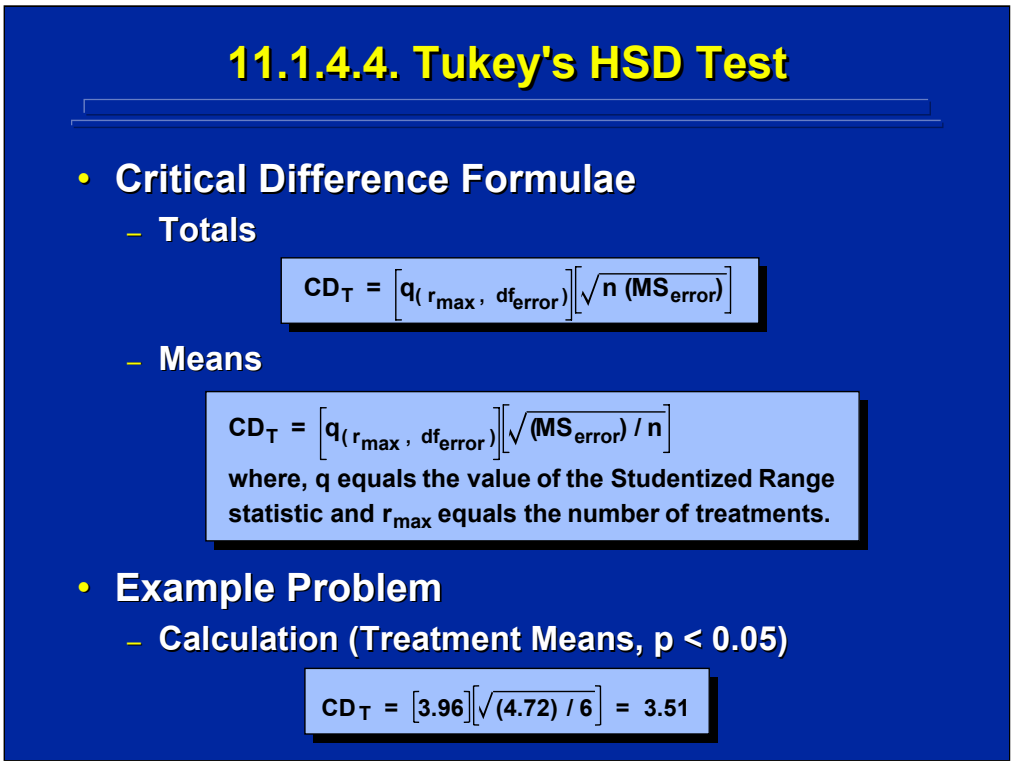
Based on the critical difference between means formula presented on this slide for the Bonferonni t test, a difference of 3.66 between any pair of means is significant. Consequently, only the difference between means $a_2$ and $a_3$ is significant at the 0.05 level.




(Click in this red rectangle to see SAS calculations for this example.)

Scheffé multiple contrast test can be used for complex comparisons as well as paired comparisons using the t sampling distribution. Consequently Type I error is distributed over a larger range of comparisons and results in a more conservative test than just considering paired comparisons.

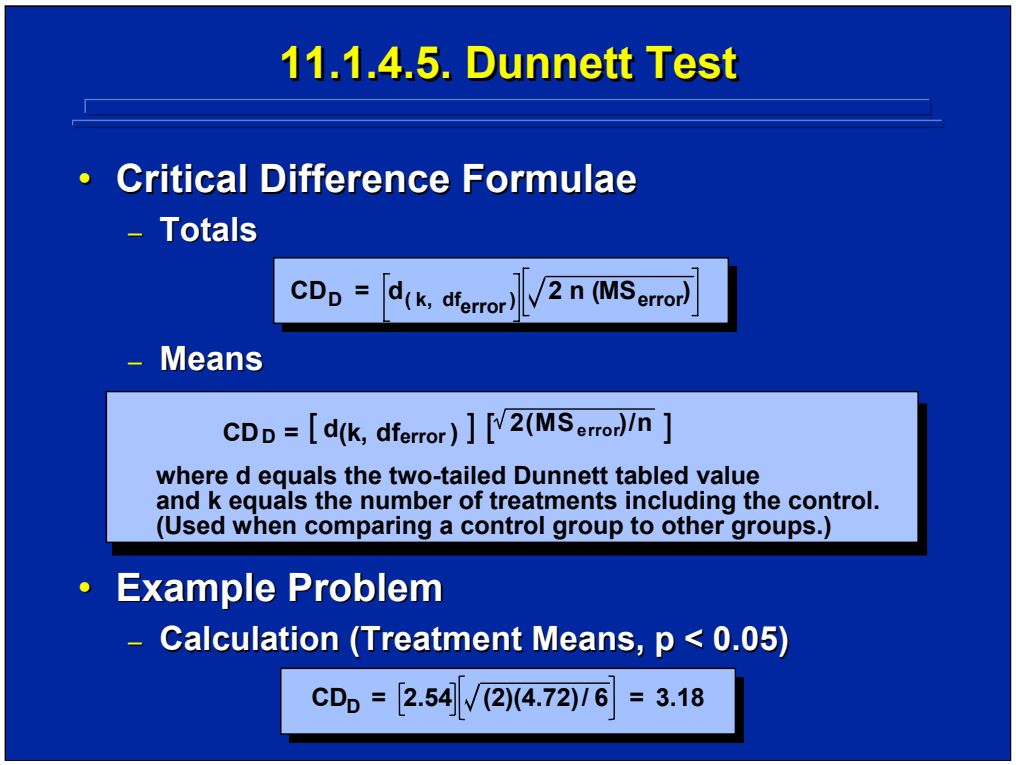
Based on the critical difference between means formula presented on this slide for the Scheffé test, a difference of 3.82 between any pair of means is significant. Consequently, only the difference between means $a_2$ and $a_3$ is significant at the 0.05 level.




(Click in this red rectangle to see SAS calculations for this example.)

Tukey's Honestly Significant Difference (HSD) test also allows simple and complex comparisons to be used, but it uses a different sampling distribution than the Scheffé test. The Tukey test uses the Studentized Range statistic, q, which is based on the maximum range, $r_{max}$, of mean differences. The value of $r_{max}$ is the total number of treatments involved the paired comparisons. The tabled value of q (i.e., $q_{(4.20)} = 3.96$) is presented in Appendix Table D.4 in Winer, et al. (1991).

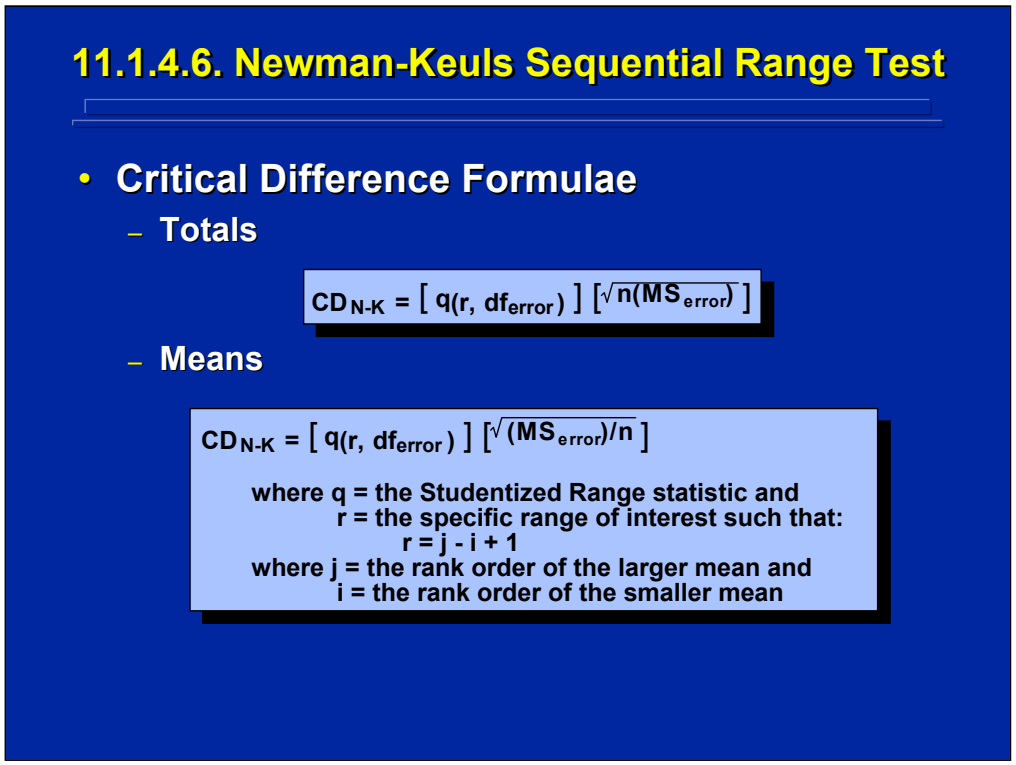
Based on the critical difference between means formula presented on this slide for Tukey's HSD test, a difference of 3.51 between any pair of means is significant. Consequently, only the difference between means $a_2$ and $a_3$ is significant at the 0.05 level.




(Click in this red rectangle to see SAS calculations for this example.)

The Dunnett test assumes one of the levels is a control condition and the other levels are experimental conditions. This test is appropriate only when there is a control group. The visual display ($a_3$) is considered the control condition in this example. Only paired comparisons of each experimental condition to the control condition are made thereby resulting in a smaller set of comparisons for distributing $\alpha$ error. These comparisons use the two-tailed Dunnett tabled value of d (i.e., $d_{(4,\ 20)} = 2.54$) found in Appendix Table D.6 in Winer et al. (1991).

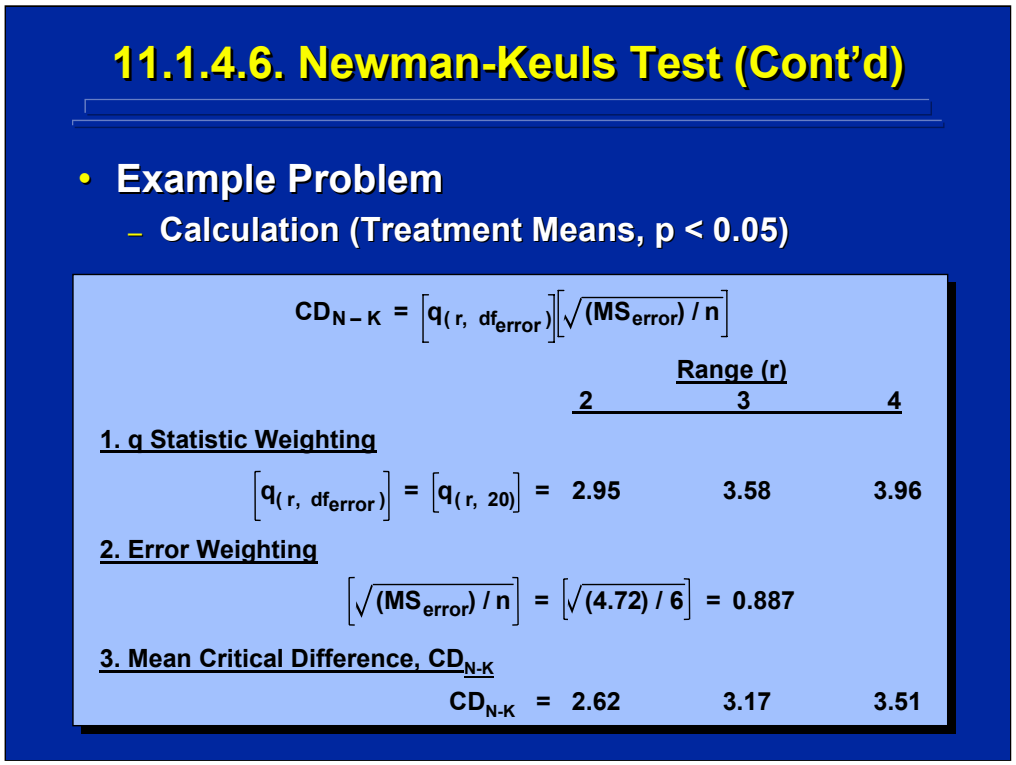
Based on the critical difference formula presented on this slide for the Dunnett test, a difference of 3.18 between any pair of means is significant. Consequently, only the difference between means $a_2$ and $a_3$ is significant at the 0.05 level assuming one of those treatments is a control condition.




(Click in this red rectangle to see SAS calculations for this example.)

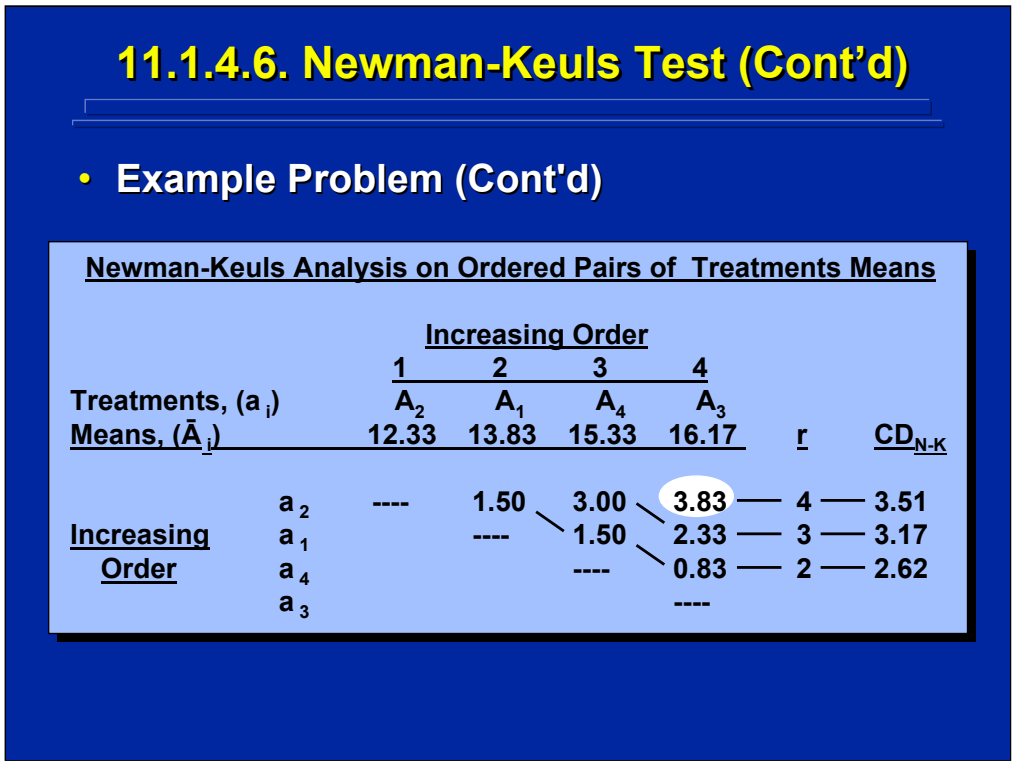
The Newman-Keuls test is a compromise paired comparison test that selectively controls for inflated Type I error by considering the ordered range of differences among treatments. It distributes the correction for $\alpha$ error depending on how far apart the means are when they are rank ordered. If the comparison consists of a pair of means that are farther apart in the rank order, there is more correction than a pair of means that are close together in the rank order.

The critical difference formulae for totals and means are shown on the slide. This test uses the Studentized Range statistic, q, tabled value presented in Appendix Table D.4 in Winer, et al. (1991). Note that the Tukey test also used the q statistic at only one value, $r_{max}$. The Newman-Keuls test, however, uses a series of values that are based on the sequential range, r, of the paired comparison in order to determine the tabled value of q. The paired difference of means closest together has a range of 2 (i.e., 2-1+1), and the paired difference of the means farthest apart in the range has a value of r. Hence the range goes from 2 to r where r equals the total number of different means in the paired comparisons.




(Click in this red rectangle to see SAS calculations for this example.)

This slide provides the calculations for the three values of critical differences between means based on q values that range from 2, 3, and 4 for all the paired comparisons among the 4 means in the example problem. Note that the weighting of error, 0.887, is constant at each range, r. The various critical differences between means are obtained by multiplying each of the q values listed in Appendix Table D.4 in Winer et al. (1991) for ranges 2, 3, and 4 (i.e., 2.95, 3.58. and 3.96, respectively) by 0.887. The resulting three critical differences between pairs of means, $CD_{N-K}$, are 2.62, 3.17, and 3.51 for this example on the bottom line of this slide.




(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the increasing rank order of the four means in the example problem. The resulting table of mean differences shows each of the resulting six pairs of differences according to its sequential range, r. Note that three paired differences are at range 2 (i.e., 1.50, 1.50, and 0.83, two paired differences at range 3 (i.e., 3.00, 2.33), and one paired difference at range 4 (i.e., 3.83). The experimenter must evaluate each difference in terms of its range using the appropriate $CD_{N-K}$ value.

In the Newman-Keuls test, the experimenter compares the resulting difference in each pair of means to the critical difference calculated for its particular range in order to determine significance. The three critical differences, $CD_{N-K}$, for this example problem as calculated on the previous slide are listed in the right most column of this slide. As shown in the ellipse on this slide, only the mean difference of 3.83 at r = 4 is greater than its respective $CD_{N-K}$ of 3.51. The mean differences of 3.00 and 2.33 are not greater then their appropriate critical difference of 3.17 nor are 1.50, 1.50, and 0.83 greater than the critical difference 2.62. Consequently, only the difference between means $a_2$ and $a_3$ is significant at the 0.05 level according to the Newman-Keuls test.

## 11.1.4.7. Choice of Procedure

**Comparison of Critical Differences of Unplanned, Paired-Comparison Procedures**

| Procedure | Range (r) 2 | 3 | 4 |
|---|---|---|---|
| LSD (Multiple t Test) | 2.62 | 2.62 | 2.62 |
| Newman-Keuls Test | 2.62 | 3.17 | 3.51 |
| Dunnett Test | 3.18 | 3.18 | 3.18 |
| Tukey's HSD Test | 3.51 | 3.51 | 3.51 |
| Bonferroni t Test | 3.67 | 3.67 | 3.67 |
| Scheffé Test | 3.82 | 3.82 | 3.82 |

This slide compares the critical differences needed to obtain significance between pairs of means in the example problem over each of the three ranges for each of the paired comparison procedures described in this subsection. Note that the Newman-Keuls procedure is the only alternative where the critical difference changes based on range.

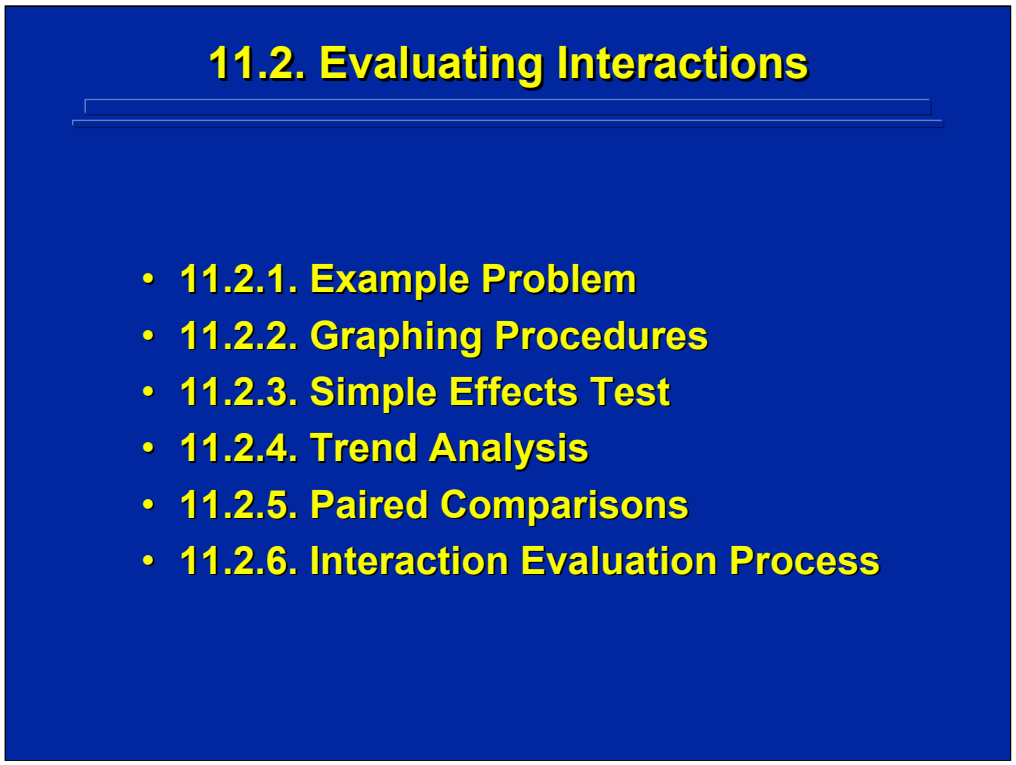
A test with a higher critical difference is more stringent in obtaining statistical significance than a test with a lower critical difference. Consequently, the Scheffé test, which distributes $\alpha$ error across all possible simple and complex comparisons, is the most stringent, and the LSD test which makes no correction for inflated $\alpha$ error is the least stringent in controlling for an inflated Type I error. The Newman-Keuls test is equivalent to the LSD test for paired comparisons at range 2 and is equivalent to Tukey's HSD test at range 4.

## 11.1.4.7. Choice of Procedure (Cont'd)

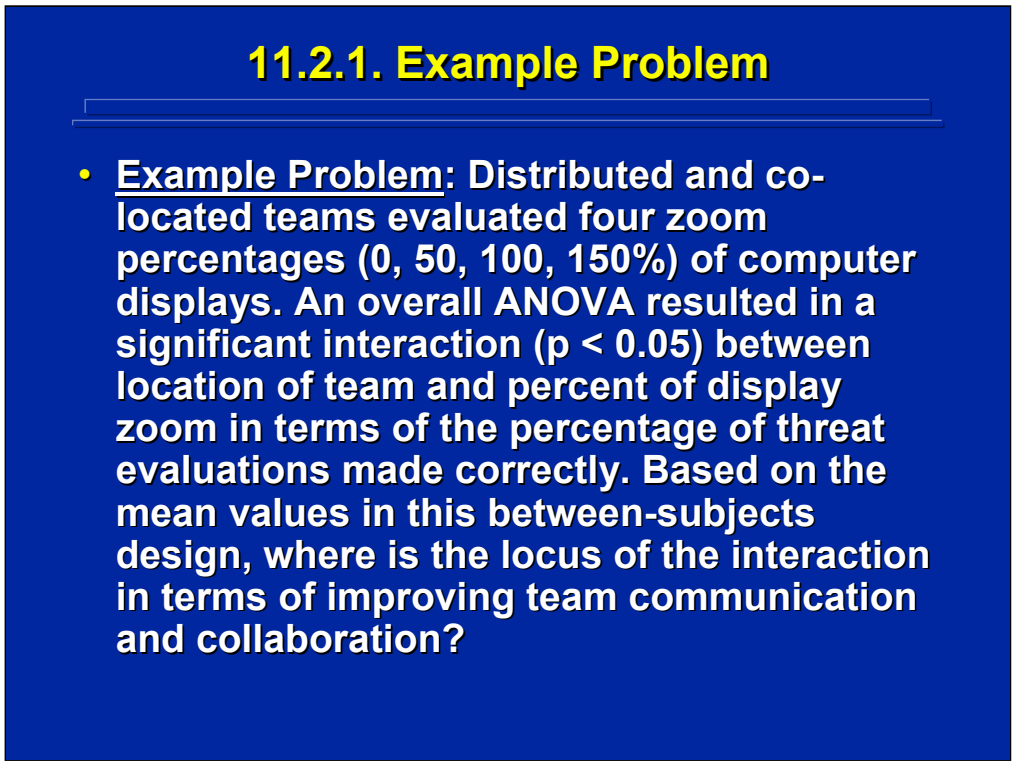
- **Considerations**
  - Type of Comparison
  - Inflation of $\alpha$ Error
  - Rationale of Procedure
  - No Single Procedure is Appropriate for All Comparisons
- **Choice of Multiple Comparison Procedure**
  - No Correction - LSD
  - Unplanned Paired Comparisons Correction - Newman-Keuls or Bonferroni t
  - Comparison to Control Condition - Dunnett
  - Complex Comparisons - Scheffé or Tukey's HSD

The experimenter needs to consider the type of comparison, the control for inflated Type I error, and the rationale of various test alternatives before making a decision as to which analytical procedure to use when making comparisons. Consequently, no single procedure is appropriate for all comparisons, and the experimenter needs to understand the available alternatives.

If the experimenter is interested in finding all possible paired comparisons that may exist in a significant ANOVA main effect or interaction and is not concerned with inflated $\alpha$ error due to the overall test of significance, then an LSD or multiple t-tests can be conducted. If control for Type I error inflation is a concern, then the Newman-Keuls and the Bonferroni t tests are appropriate for unplanned, paired comparisons. The Bonferroni t test is more stringent by making one overall correction across all comparisons, whereas the Newman-Keuls test distributes stringency depending on the range of paired differences. If comparisons are only made between a control condition and experimental conditions, then the Dunnett test is appropriate. If both simple and complex comparisons are being conducted both the Scheffé and Tukey's HSD tests are appropriate.

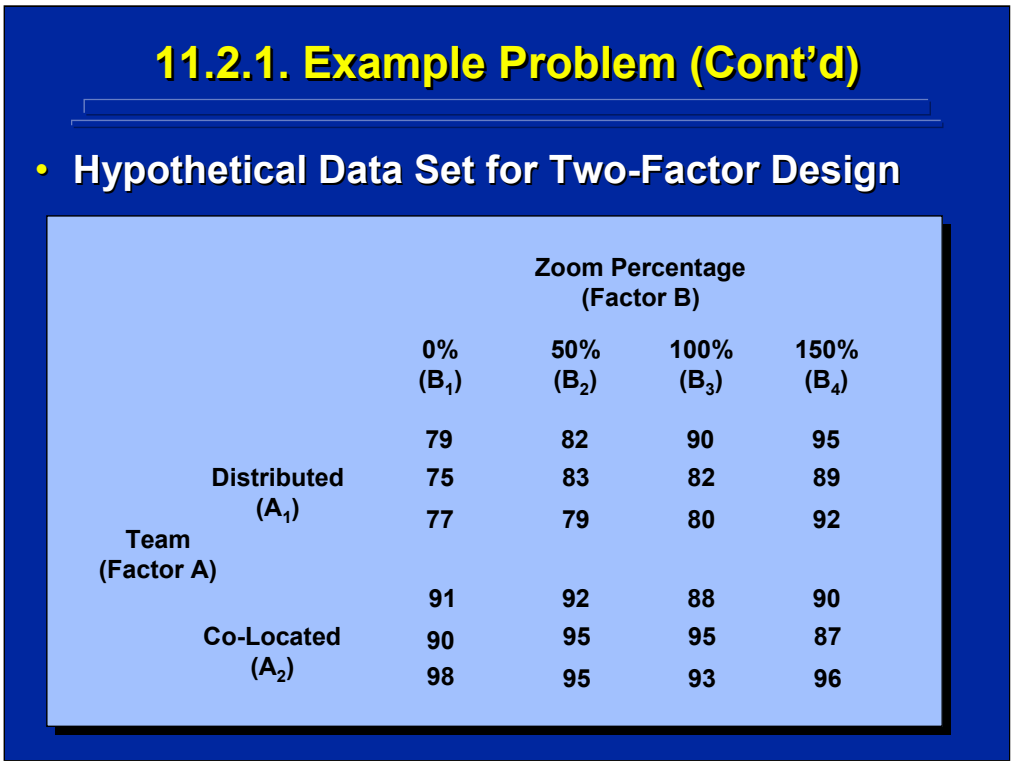



Isolating a significant interaction in ANOVA also requires post hoc analysis. An example problem is provided to demonstrate these analysis alternatives. Both graphical and analytical procedures are appropriate. Three major computational procedures involving simple effects tests, trend analysis, and paired comparisons are described in this subsection.




(Click in this red rectangle to see SAS calculations for this example.)

This slide describes a 2x4 between-subjects design that resulted in a significant interaction. Improved percent of threat evaluations resulted as a function of the interaction between Location of Teams and Percent Zoom of a computer information display used to improve team coordination. Based on these results, the experimenter is interested in determining the exact effect of the significant interaction.




(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the data set for for the 2x4 example problem described on the previous slide. Factor A has two levels of team location and Factor B has 4 levels of percent zoom of the computer displays. Since this is a between-subjects factorial design, a total of 24 teams of threat evaluators are used across the 8 cells comprising the AxB interaction in the experiment.

## 11.2.1. Example Problem (Cont'd)

- **AxB Interaction of a Between-Subjects Design**

**ANOVA Summary Table**

| Source | df | SS | MS | F |
|---|---|---|---|---|
| A | 1 | 477.04 | 477.04 | 37.66* |
| B | 3 | 128.12 | 42.71 | 3.37 |
| AxB | 3 | 245.13 | 81.71 | 6.45* |
| S/AB | 16 | 202.67 | 12.67 | |
| Total | 23 | 1052.96 | | |

* p < .01

(Click in this red rectangle to see SAS calculations for this example.)

The ANOVA Summary Table for the 2x4 example problem is presented on this slide in general form. Factor A (Location of Teams) has two levels, distributed and co-located teams. Factor B (Percent Zoom of Computer Display) has four levels, 0, 50, 100, and 150% zoom. The error term for all F-tests in this two-way, between-subjects design is $MS_{S/AB}$. Both Factor A and the AxB interaction are significant. Further analyses are needed to find the locus of the AxB interaction.

## 11.2.1. Example Problem (Cont'd)

- AxB Interaction
  - n = 3 (Note: 3 x 8 = 24 observations)
  - $MS_{error} = MS_{S/AB} = 12.67$
  - $df_{error} = df_{S/AB} = 16$
  - $AB_{ij.}$ Totals

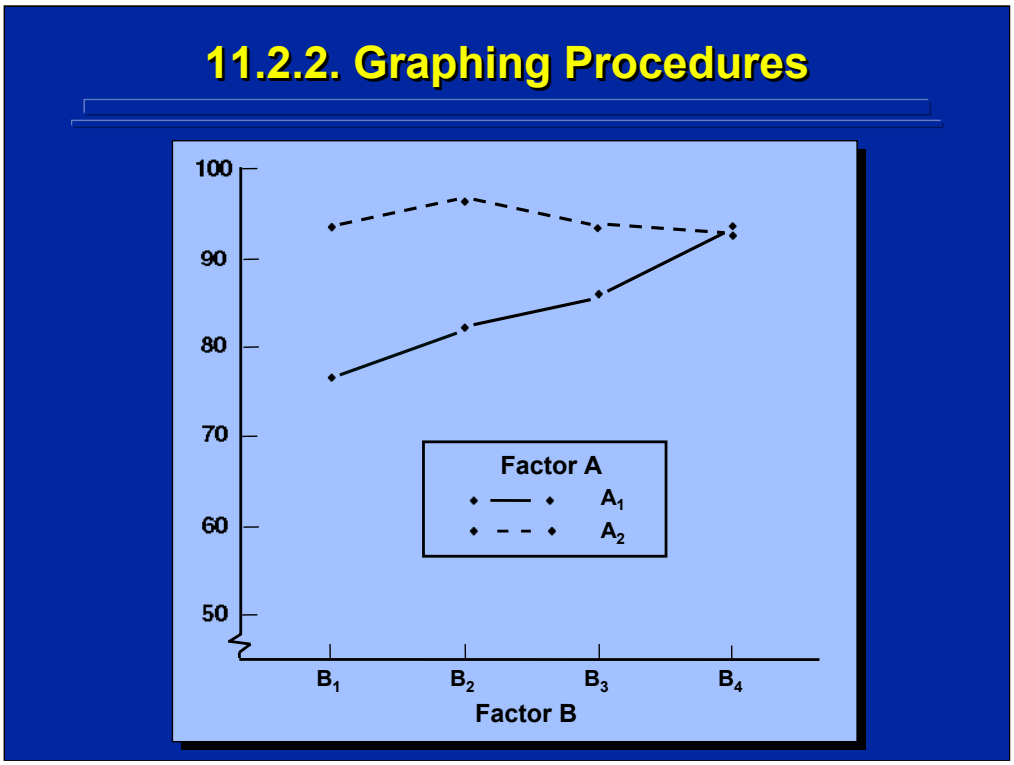
| | |
|---|---|
| $AB_{11.} = 231$ | $AB_{21.} = 279$ |
| $AB_{12.} = 244$ | $AB_{22.} = 282$ |
| $AB_{13.} = 252$ | $AB_{23.} = 276$ |
| $AB_{14.} = 276$ | $AB_{24.} = 273$ |

- Interaction = Differential Effect
- Interpretation of Interaction
  - Graphical Procedures
  - Numerical Procedures

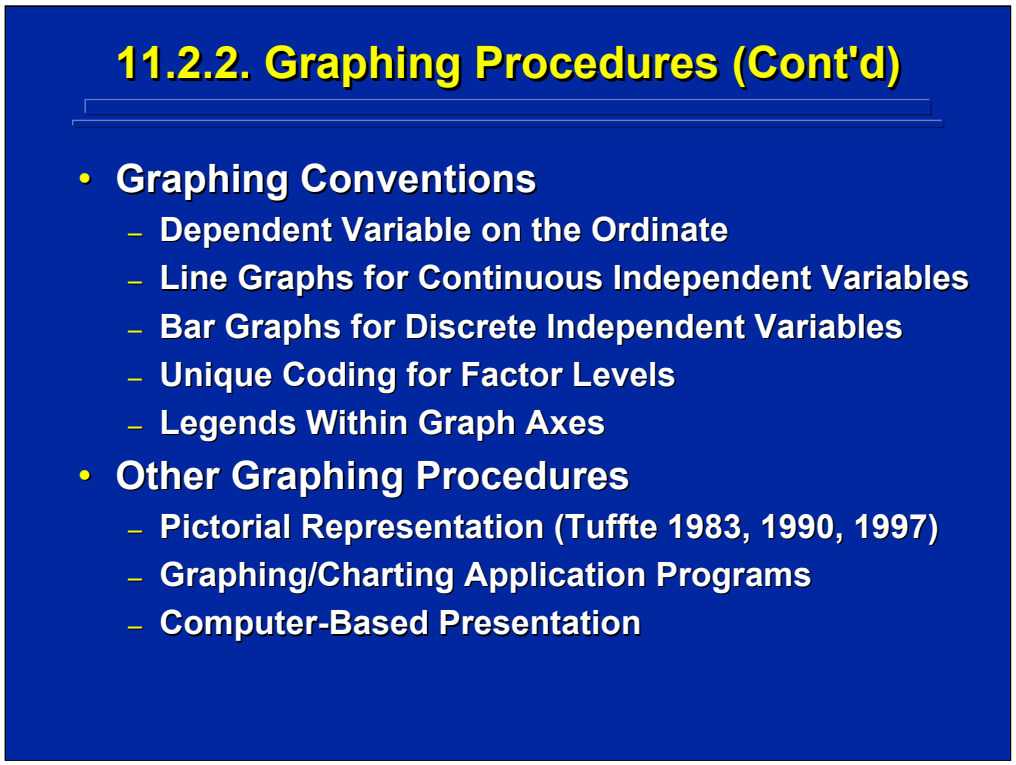
(Click in this red rectangle to see SAS calculations for this example.)

If one assumes equal sample size, the sample size, n, for any interaction times the number of treatments in the interaction, t, equals the total number of observations in the experiment, N. Consequently, n in this experiment equals 3. The $MS_{Error}$ and the $df_{Error}$ for evaluating the interaction are obtained from the error term used in the overall ANOVA. They are 12.67 and 16, respectively, as presented in the previous slide.

The totals of the three scores for each of the eight treatment combinations in the AxB interaction are presented in the middle of this slide. It appears that there is an increase in scores as B changes across the first level of A and that there is relatively little change in scores as B changes across the second level of A. But, additional graphical and numerical analyses are needed to confirm this interaction effect.

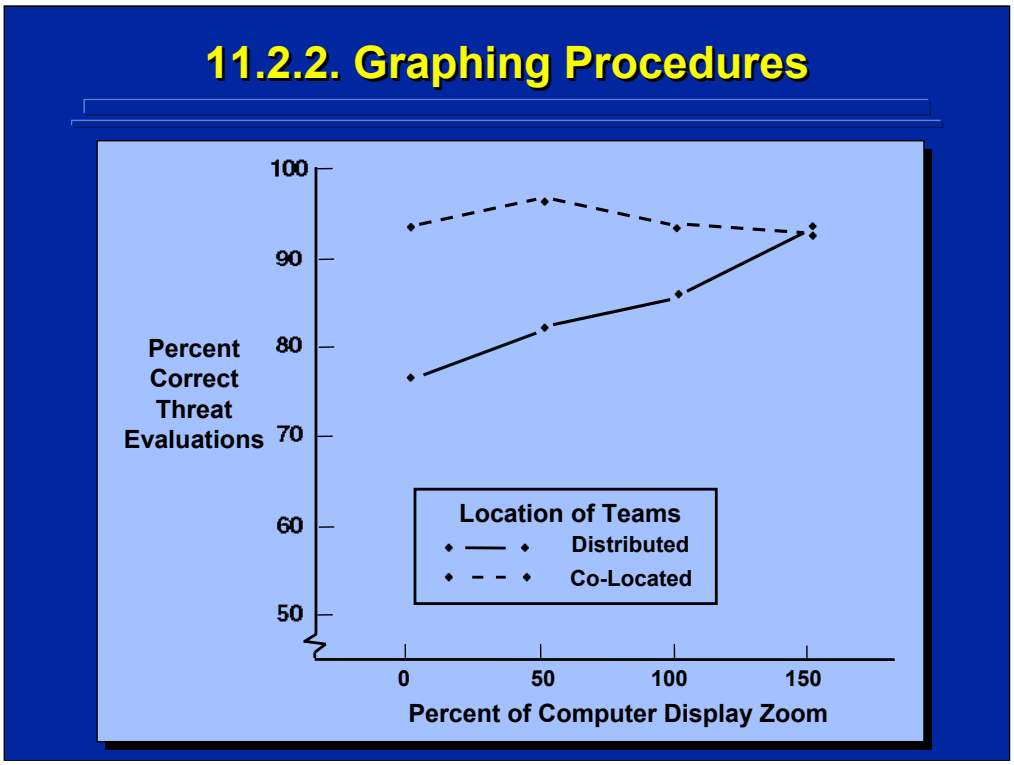



Means of the eight treatment conditions totals of the AxB interaction data listed on the previous slide are presented on a graph to aid in interpretation of the interaction effect. The two levels of Factor A are plotted as separate lines across the four levels of Factor B. The plot is presented in black and white as used in most human factors publications. The broken vertical line on the ordinate indicates that scores could be below 50. If the entire scale from 0 to 100 was shown on the graph, the interaction difference would appear smaller, but a great deal of blank space would appear in the graph.



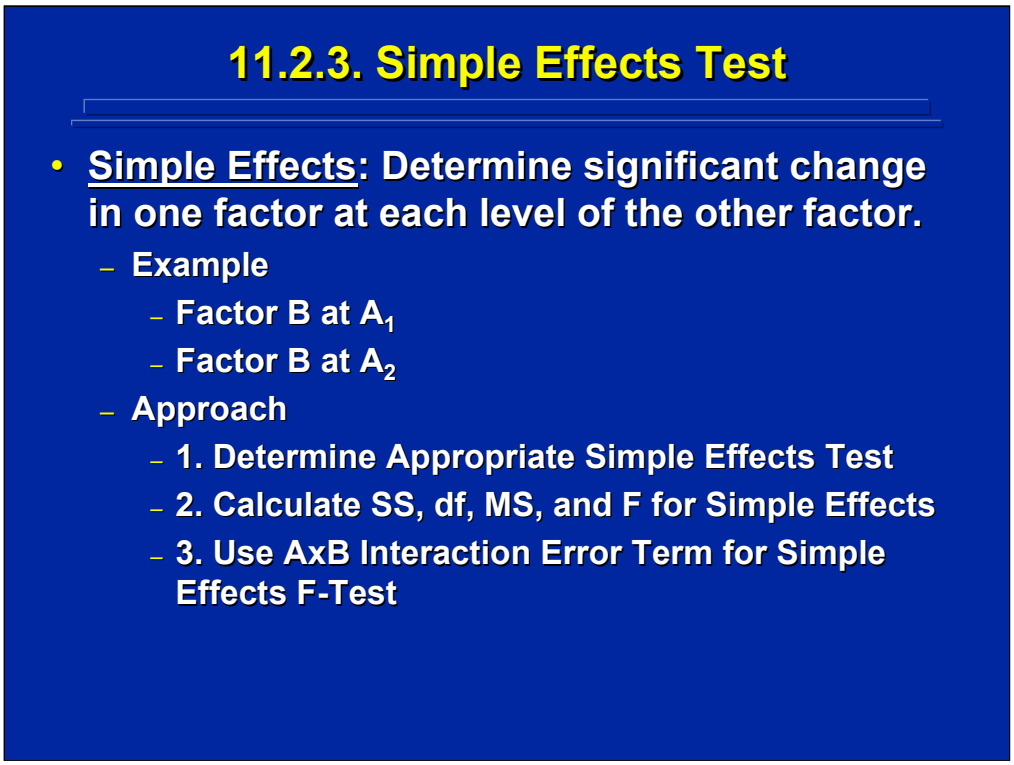

Some graphing conventions are listed on this slide pertain to two-dimensional, black and white slides often used in human factors and ergonomics research. The dependent variable is listed in the ordinate, and one independent variable is listed on the abscissa. The other independent variable(s) in the interaction are plotted in the graph, and the levels are represented either as lines are bars. Line graphs are used for continuous variables, whereas bar graphs are used for discrete variables. A unique coding such as solid, dashed, and dotted lines are used to designate factor levels. The legend defining these various levels should remain within the graph.

Other graphing procedures like perspective bar graphs and color coding are often used. Tuffte (1983, 1990, and 1997) provides a variety of innovative pictorial and graphical representations of data to improve interpretation. Modern computer graphing and plotting techniques provide many alternatives to the researcher to improve communication of the interaction effect to the reader.




This slide shows a re-plot of the AxB interaction using the standard graphing procedures summarized in the previous slide and stating the factors and levels in terms of the actual experiment. The dependent variable used in the experiment is listed as the ordinate. Line graphs are used because percent of display zoom is a continuous variable. The two levels of Location of Teams are plotted as dashed and solid lines and defined in the figure legend contained within the graph boundary.

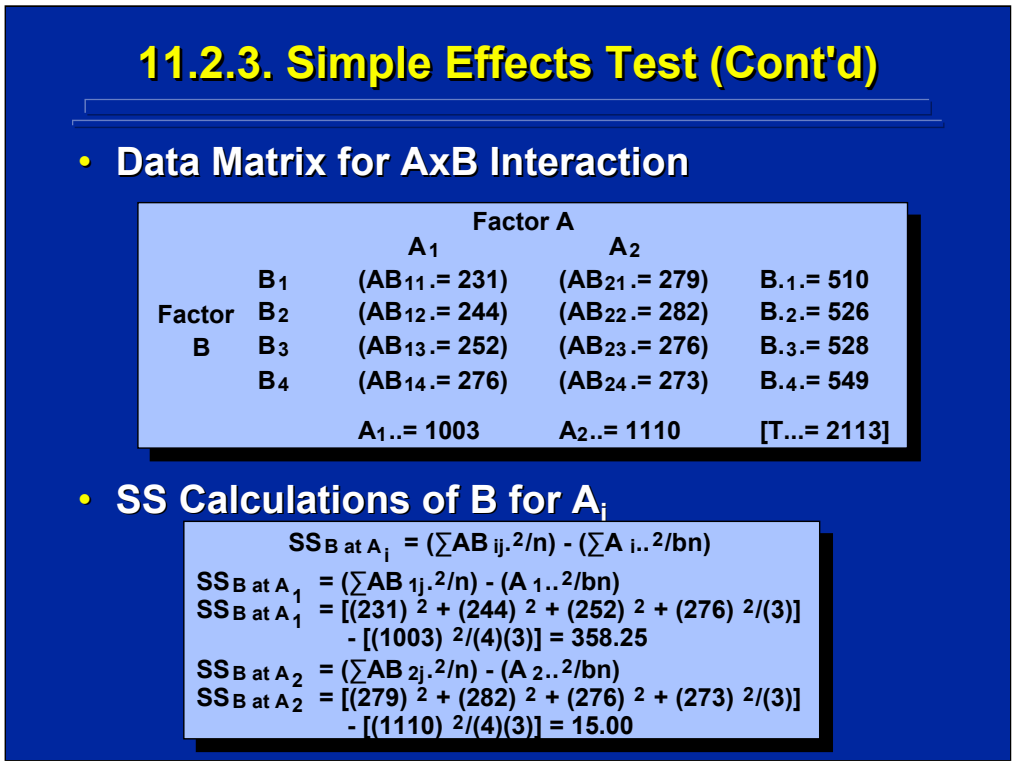
Statistically significant differences cannot be inferred from the graph directly. Additional analytical procedures are necessary to isolate all the significant effects. First, one might conduct a simple effects test that restricts significant differences across the four levels of Display Zoom (Factor B) to just distributed teams ($A_1$) and then to just co-located teams ($A_2$). Second, there seems to be a linear increase across display zoom in distributed teams that can be verified by a trend analysis. Finally, the difference between distributed and co-located team performance at the 0% zoom display level can be tested for statistical significance using paired comparisons. Each of these analytical techniques is described separately.




(Click in this red rectangle to see SAS calculations for this example.)

One way to isolate overall interaction effects is to test for significant differences across one factor at a particular level of the other factor. This is referred to as a simple effects test. The two examples presented on the slide are, first, the changes across levels of Factor B at the $A_1$ level of Factor A and, second, the changes across levels of Factor B at the $A_2$ level of Factor A.

The three general steps involved in a simple-effects test are listed on the bottom of this slide. First, the experimenter must determine whether the simple effect is tested across Factor A at each level of Factor B or vice versa. Second, the calculations of the appropriate simple effects tests are made. Third, the overall error term of the interaction is used as a pooled error term for all simple effects tests.




(Click in this red rectangle to see SAS calculations for this example.)

The data for the AxB interaction in the example problem are shown in the top portion of this slide. The bottom portion of the slide lists the general formulae for calculating the SS of a simple effect. This is simply using data at each level of Factor A independently to compute a $SS_B$; i.e., compute $SS_B$ at level $A_1$ and ignore data at level $A_2$, then repeat using data at only level $A_2$ and ignore data at level $A_1$. Using this formula for calculation, the $SS_{B \text{ for } A1}$ = 358.25 and the $SS_{B \text{ for } A2}$ = 15.00 for the example data.

## 11.2.3. Simple Effects Test (Cont'd)

**ANOVA Summary Table For Simple Effects**

| Source | df | SS | MS | F |
|---|---|---|---|---|
| B at Level $A_1$ | 3 | 358.25 | 119.42 | 9.42* |
| B at Level $A_2$ | 3 | 15.00 | 5.00 | <1.00 |
| S/AB | 16 | 202.67 | 12.67 | |

*$p < 0.05$

Note: $SS_{B \text{ at } A1} + SS_{B \text{ at } A2} = SS_{B/A} = SS_B + SS_{AxB} = 373.25$

(Click in this red rectangle to see SAS calculations for this example.)

The ANOVA Summary Table shown on this slide provides the F-tests based on the error term used for the overall AxB interaction test of significance (i.e., pooled error term). The simple-effect test of Factor B for $a_2$ shows that none of these paired differences are significant at the 0.05 level. Only the simple effect of Factor B for $A_1$ is significant. This means that at least one pair of the four levels of B are significantly different at the $A_1$ level of Factor A, but the exact differences cannot be determined by the simple effects test. Subsequent paired comparisons of the four levels of Factor B at $A_1$ need to be conducted to isolate these differences.

In terms of the factors manipulated in the example problem, the simple-effects analysis summarized on this slide means that computer display zoom fails to affect co-located team performance. However, display zoom does significantly affect distributed team performance at the 0.05 level. The last line of this slide notes that the total of the SS for the two simple-effects tests (373.25) is equal to $SS_B$ plus $SS_{AxB}$ in the Summary Table of the overall ANOVA on a previous slide in this example.

## 11.2.4. Trend Analysis

- **Trends**: The nature of the relationship between treatment condition magnitudes, t, and the dependent variable magnitudes, Y.
  - Assumes Quantitative Independent Variables
  - Equally Spaced Factor Levels, k
  - Relationship Expressed as Polynomials
    - $Y = b_0 + b_1t + b_2t^2 + b_3t^3 + ... + b_{k-1}t^{k-1}$
      - $b_0$ = Constant
      - $b_1t$ = Linear Component
      - $b_2t^2$ = Quadratic Component
      - $b_3t^3$ = Cubic Component
      - $b_{k-1}t^{k-1}$ = k-1 Component

(Click in this red rectangle to see SAS calculations for this example.)

Trend analysis evaluates the quantitative relationship among treatment means. It is a special case analysis for quantitative factors that are usually manipulated as equally spaced levels in the experiment. Myers (1979, pp. 441-445), however, shows a computational procedure for conducting a trend analysis when the levels are not equally spaced and transformations to obtain equal spacing are not appropriate. The quantitative relationship of the dependent variable, Y, is also expressed as a weighted orthogonal polynomial of various linear and curvilinear components, t. Trend analysis can be used to interpret both main effects and interactions.

In the example problem of the AxB interaction, the simple-effects test demonstrated that computer display zoom significantly affected threat evaluation performance of distributed teams. The graph of this interaction seems to show a linear increase in performance as the four quantitative and equally spaced levels of computer display zoom decrease (i.e., 0, 50, 100, and 150% zoom). A subsequent trend analysis on this simple effect would confirm if there is a significant linear increasing trend.

## 11.2.4. Trend Analysis (Cont'd)

- Order of Trend
  - Highest Order Trend One Less Than Number of Levels
  - Linear or Quadratic Trends Fit Most Human Factors Data
- Numerical Value of a Trend Comparison for Totals, $T_j$
  - $C_{Trend} = \Sigma c_{ij} T_j \quad (\Sigma c_{ij} = 0, \ \Sigma c_{ij} c_{i'j} = 0)$
  - Where $c_{ij}$ = Tabled Orthogonal Polynomial Coefficients
- Orthogonal Polynomials Have Independent Terms
  - $SS_{Trend} = SS_{Linear} + SS_{Quadratic} + SS_{Cubic} + \ldots + SS_{k-1}$
- Sum of Squares Calculations Based on Treatment Totals

$$SS_{Trends} = \frac{C_i^2}{n\Sigma c_{ij}^2}$$

- $F_{Observed} = SS_{Trend} / MS_{Error}$
- $F_{Tabled} = 1, df_{Error}$

(Click in this red rectangle to see SAS calculations for this example.)

In general, the highest trend that can be evaluated is one less than the number of equally-spaced levels of the quantitative factor. Linear and quadratic trends tend to fit most human factors datasets thereby requiring a minimum of three levels of a factor.

Orthogonal polynomials are used in trend comparisons to keep the SS associated with each trend additive. The sum of the weights and the sum of the cross products of the weights must equal zero to keep the trend effects orthogonal and independent of the grand mean. Orthogonal coefficient weights used to test for linear, quadratic, cubic, etc. trends are provided in Table D.10 in Winer et al. (1991). The numerical value of a comparison of trends is calculated using the sum of the appropriate orthogonal polynomial coefficient weighting times the total score for all observations for each treatment level. The formula for calculating the SS of a trend is shown at the bottom of this slide. The SS of a trend is divided by $MS_{Error}$ to yield an $F_{Observed}$ value which is compared to the $F_{Tabled}$ value to test for the trend effect.

## 11.2.4. Trend Analysis (Cont'd)

- **Calculation of Trends for $Bxa_1$ Simple Effect**

**Orthogonal Trend Coefficients**

| | Treatment Totals | | | |
|---|---|---|---|---|
| $T_j$ = | $B_1$ (231) | $B_2$ (244) | $B_3$ (252) | $B_4$ (276) |
| **Trend** | | | | |
| Linear | -3 | -1 | 1 | 3 |
| Quadratic | 1 | -1 | -1 | 1 |
| Cubic | -1 | 3 | -3 | 1 |

**Linear Trend of B at $A_1$**

$C^2_{Linear} = [(-3)(231) + (-1)(244) + (1)(252) + (3)(276)]^2 = 20{,}449$

$n\Sigma c_{1j}^2 = 3[(-3)^2 + (-1)^2 + (1)^2 + (3)^2] = 60$

$SS_{Linear} = C^2_{Linear} / n\Sigma c_{ij}^2 = 20{,}449/60 = 340.82$

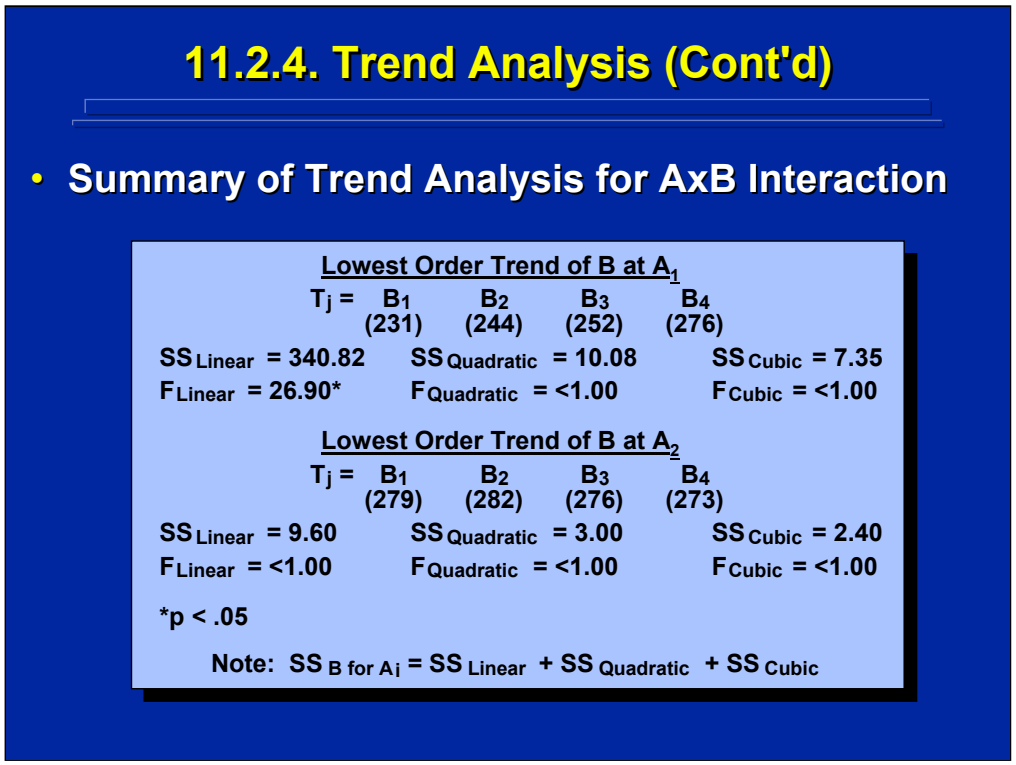
$F_{Linear} = SS_{Linear} / MS_{Error} = 340.82/12.67 = 26.90$

$F_{Tabled} = (1, 16) = 4.49$

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the linear trend analysis on the significant simple effect of the AxB interaction example. This trend analysis accesses the linear decrease across Factor B at the first level of Factor A. Since there are four levels of Factor B, linear, quadratic, and cubic trends can be accessed across the levels of Factor B. The orthogonal polynomial weighting coefficients from Table D.10 of Winer et al. (1991) are shown on the slide.

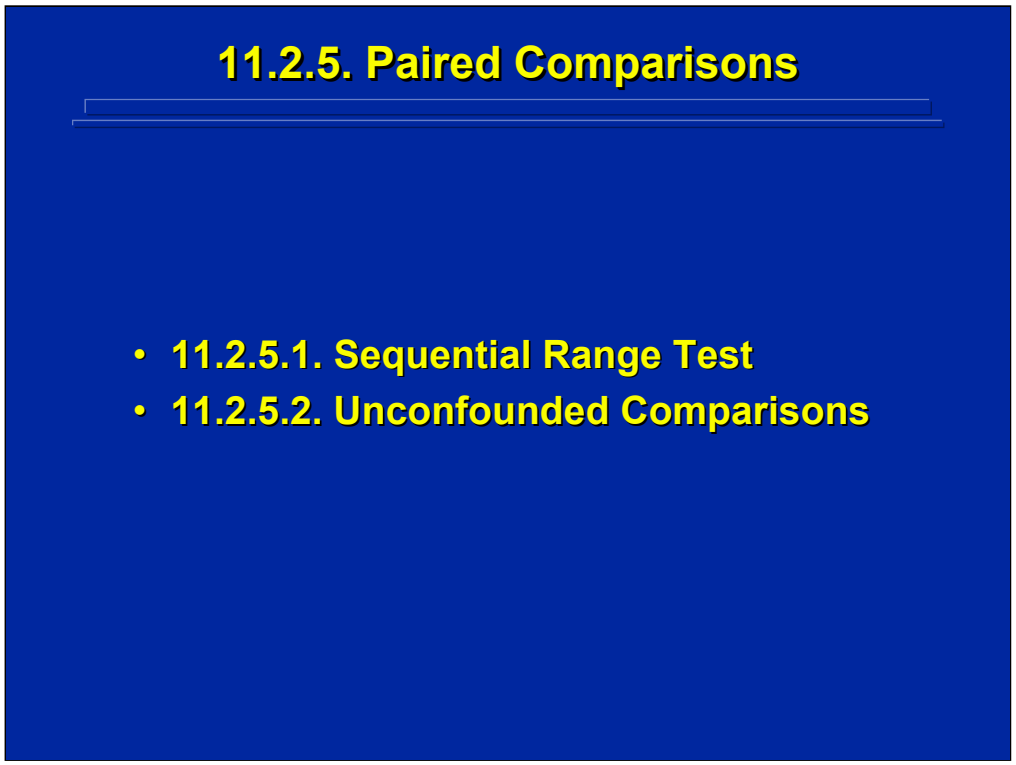
The bottom portion of this slide shows the linear trend analysis of the simple effect. The SS for this linear trend is 340.82 yielding an observed F ratio equal to 26.90 for the linear trend. Since the observed F value is greater than the tabled value of $F_{(1,16)} = 4.49$, there is a significant linear trend. Consequently, distributed teams demonstrated a linear increase in threat evaluation performance as computer display zoom decreased.



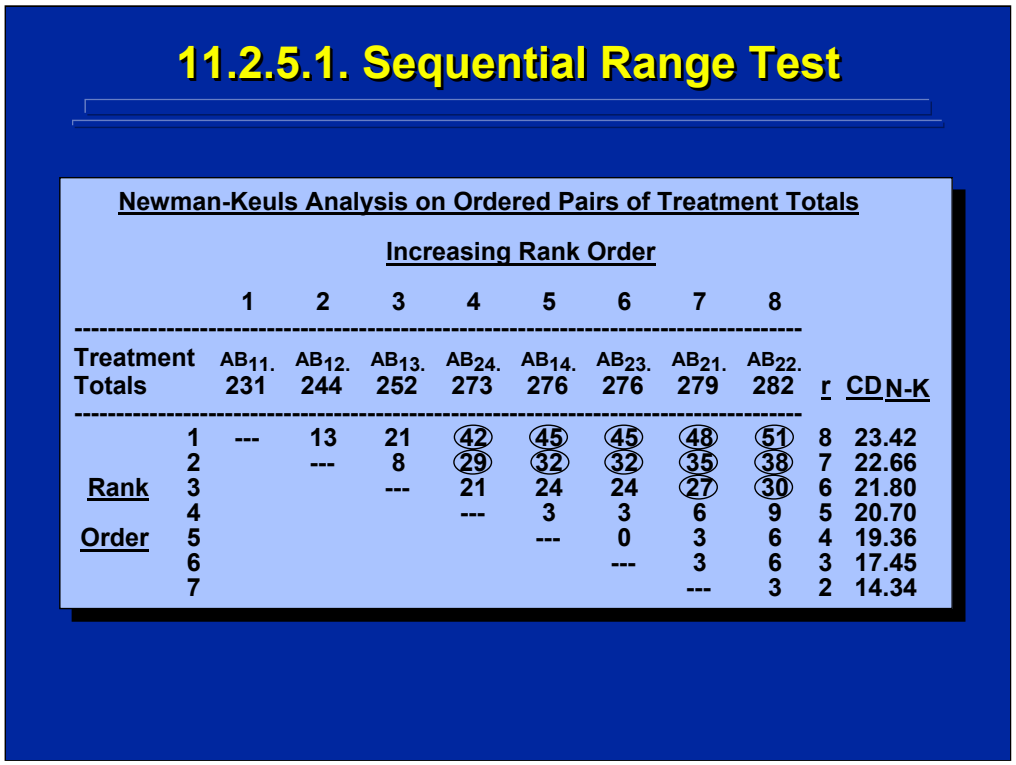

(Click in this red rectangle to see SAS calculations for this example.)

This slide summarizes the six possible trend analyses that can be conducted on the simple effects of the example AxB interaction. Note that only the linear trend of Factor B for the $A_1$ level of Factor A as shown on the previous slide is significant at the 0.05 level of significance. Since the trends are orthogonal, the sum of the $SS_{Linear}$, $SS_{Quadratic}$, and $SS_{Cubic}$ is equal to the SS of the simple-effects test.

Trend analysis is useful for the researcher to describe the quantitative relationship of interaction simple effects and main effects in ANOVA that involve equally spaced levels of quantitative factors. Paired comparisons of treatment levels, however, are still needed to determine any significant differences among the treatment levels.

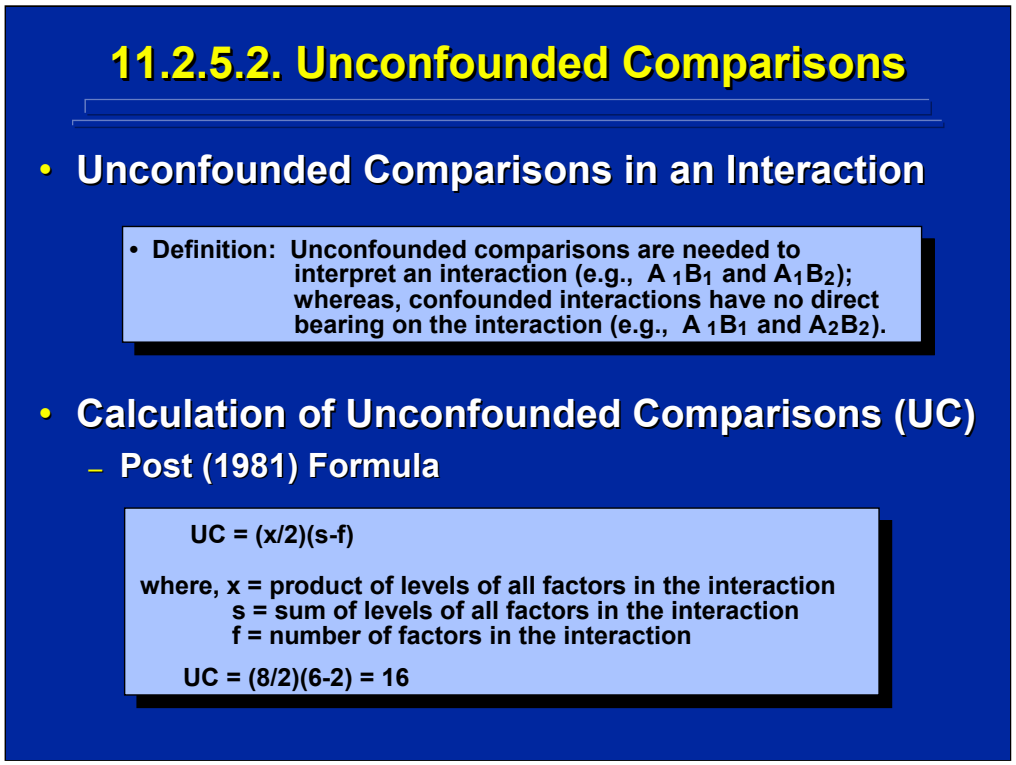



Unplanned paired comparisons are most often used to isolate interaction effects. Paired comparisons can be used to analyze simple effects of interactions, and they can be used directly on the overall interaction treatments. This subsection demonstrates the use of the Newman-Keuls test on all paired comparisons present among the interaction treatments as well as the use of unconfounded comparisons that pertain only to interaction effects.




(Click in this red rectangle to see SAS calculations for this example.)

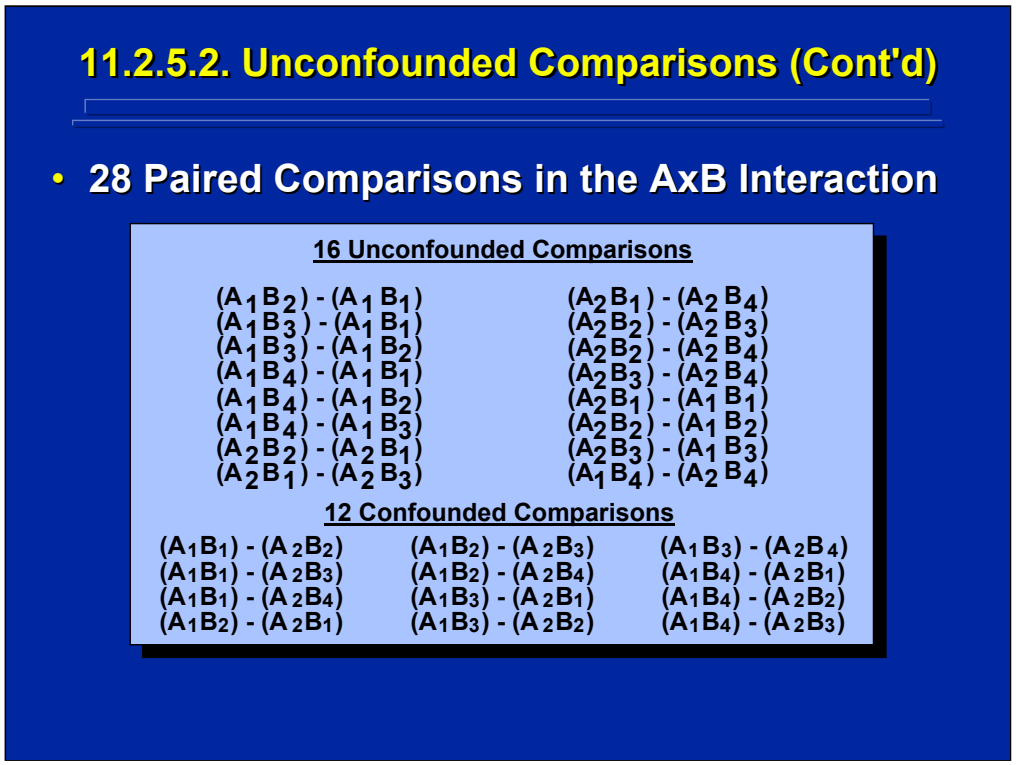
This slide shows the results of a Newman-Keuls Sequential Range test conducted on the 28 possible paired comparisons of the eight treatment totals in the 2x4 interaction of the example problem. Every difference that is circled on the slide is a significant difference at the 0.05 levels of statistical significance. Not all of these differences directly relate to the interpretation of the interaction. Consequently, the experimenter must refer to the graph of the interaction to determine which differences are useful in interpreting the interaction. For example, the type of team location has no effect on threat evaluation at 150% computer display zoom (i.e. $A_1B_4 – A_2B_4$), but there is a significant difference in threat evaluation preference between co-located and distributed teams using the 0% zoom level of computer displays (i.e., $A_1B_1 – A_2B_1$).




(Click in this red rectangle to see SAS calculations for this example.)

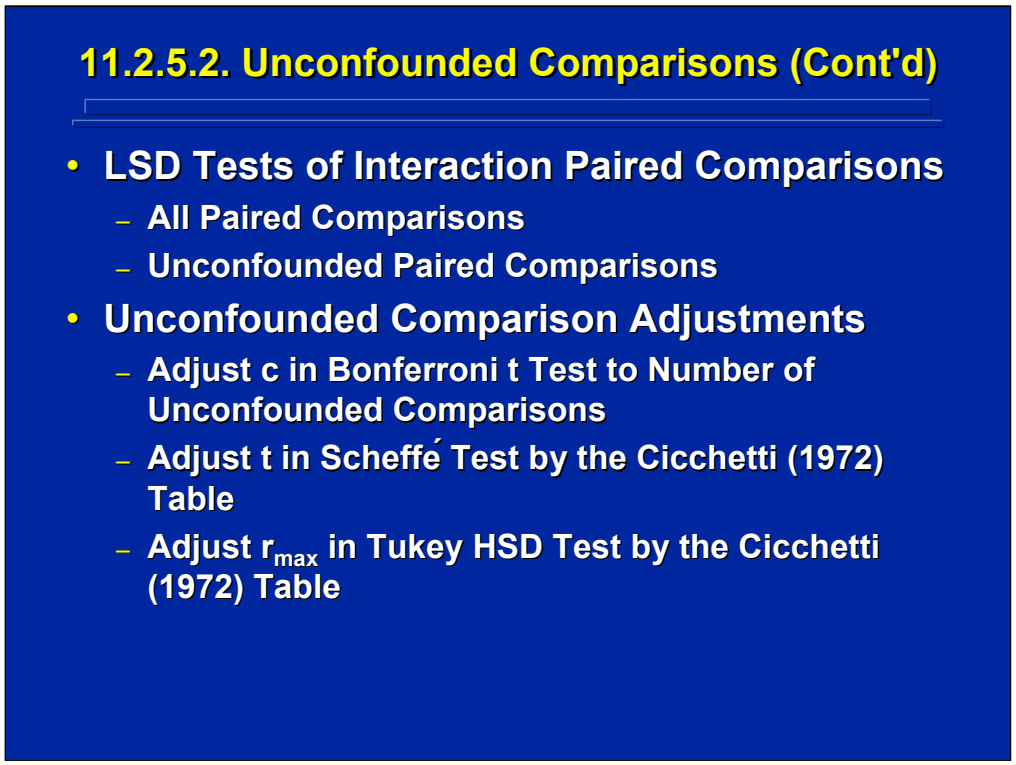
The subset of paired comparisons that relate directly to interactions is called unconfounded comparisons. These unconfounded comparisons always have one level of one of the factors in common across the paired comparison (e.g., $A_1B_1 – A_1B_2$). Paired comparisons that have different levels of factors in the paired comparison (e.g., $A_1B_1 – A_2B_2$) have no bearing on the differential effect of the interaction and are called confounded comparisons.

By using the Post (1981) formula shown on this slide, only 16 of the 28 paired comparison conducted in the previous Newman-Keuls test are unconfounded comparisons. Consequently, some researchers feel that a Newman-Keuls test may not be appropriate for the post hoc analysis of an interaction because many of the paired comparisons are confounded. In fact, the SAS computerized procedure does not allow the Newman-Keuls test for post hoc analysis of interactions, but Slater and Williges (2006) demonstrate a SAS procedure for conducting a Newman-Kuels analysis on an interaction if the experimenter chooses to do so.




(Click in this red rectangle to see SAS calculations for this example.)

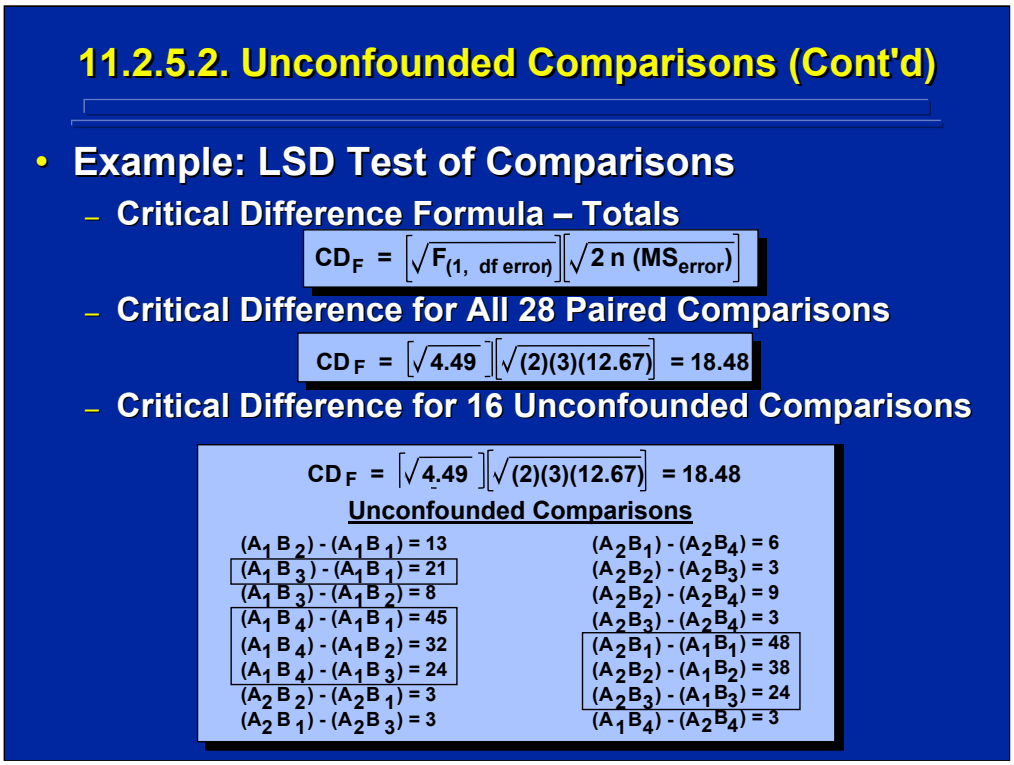
This slide lists the 16 unconfounded and 12 confounded paired comparisons of the AxB interaction in the example problem. The experimenter uses only the 16 unconfounded paired comparisons in interpreting the interaction.




(Click in this red rectangle to see SAS calculations for this example.)

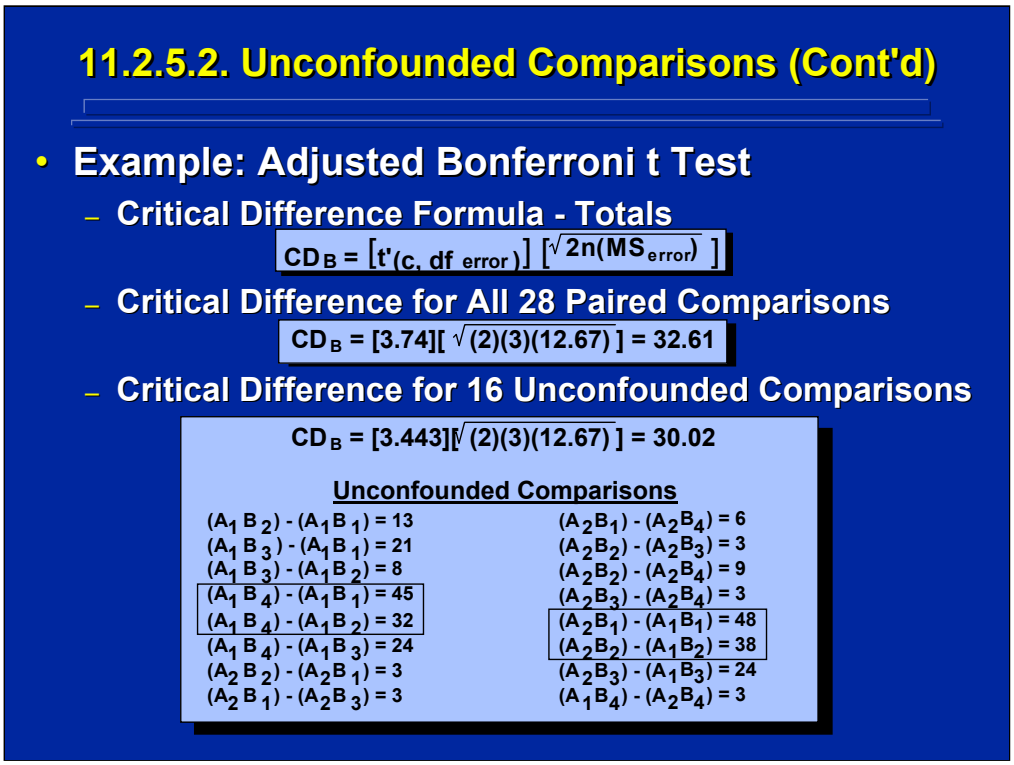
Most overall post hoc tests of paired comparisons consider both confounded and unconfounded paired comparisons when evaluating interactions as demonstrated with the Newman-Keuls test of the AxB interaction example. Such tests over control for inflated $\alpha$ error when confounded comparisons are included. The LSD test, however, makes no correction for inflated $\alpha$ error on either all paired comparisons or unconfounded comparisons involved in the interaction.

Some post hoc paired-comparison tests, however, can be adjusted for unconfounded comparisons. For example, the c used in the Bonferroni t Test could equal the number of unconfounded comparisons not the number of all possible paired comparisons in the interaction. In addition, the Scheffé Test and the Tukey HSD Test can be adjusted for unconfounded comparisons using the Cicchetti (1972) table.




(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the critical difference required between treatment totals to obtain a significant difference between treatment pairs in the AxB interaction example when using the LSD test. Note that the critical difference (18.48) is the same for all 28 paired comparisons and the 16 unconfounded comparisons involved in the interaction since no correction is made for inflated $\alpha$ error in the LSD procedure. Consequently, this is the least conservative test for isolating an overall significant interaction effect. The seven unconfounded paired comparisons of treatments involved in the interaction effect are boxed in the lower portion of this slide.




(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the effect of adjusting the Bonferroni t test for unconfounded comparisons on the AxB interaction example. As shown on the slide, the critical difference for all paired comparisons would be 32.61, whereas the critical difference for just unconfounded comparisons would be only 30.02. The difference between the $A_1B_4$ and $A_1B_2$ treatments (i.e., 32) would not be significant if the Bonferroni t test was not adjusted for unconfounded comparisons. The resulting four significant unconfounded comparisons are boxed in the lower portion of this slide.

Note that three unconfounded comparisons (i.e., $A_1B_3 - A_1B_1$, $A_1B_4 - A_1B_3$, and $A_2B_3 - A_1B_3$, ) found significant in the more lax LSD procedure shown on the previous slide would not be significant if the adjusted Bonferroni t test for unconfounded critical differences is used. Consequently, the experimenter must decide on the appropriate level of $\alpha$ error protection needed in isolating the interaction effect.

## 11.2.6. Interaction Evaluation Process

- **Step 1. Evaluate Interaction Graphically**
  - Observe Possibly Differential Effects
  - Plan Subsequent Post Hoc Analysis
- **Step 2. Evaluate Interaction Analytically**
  - Simple-Effects Test
  - Trend Analysis
  - Unplanned Paired Comparisons
- **Step 3. Interpret Impact of Interaction on Significant Main Effects**

Evaluation of significant interactions in ANOVA involves both graphical and analytical procedures in a three step process. In Step 1, the experimenter should begin by graphing the interaction data to observe possible differential effect of the interaction and plan analytical procedures to isolate interaction effect.

Every significant interaction requires a subsequent post hoc analysis conducted in Step 2. Several analytical procedures can be used. Some researchers first conduct a simple effects test to determine which level of one variable exhibits difference across the other variable. In cases involving equally spaced quantitative variables, trend analyses can be used to provide a quantitative interpretation of the simple effects. In most cases, however, the experimenter conducts unplanned paired comparisons to isolate the exact locus of the interaction.

In Step 3, the experimenter needs to interpret the impact of interactions on significant main effects. In the interaction example in this subsection, both Factor A and the AxB interaction were significant. One could state that co-located teams perform threat evaluation better than distributed teams, but this difference occurs when computer display zoom is less than 100% because distributed teams show a linear increase in threat evaluation as a function of increasing display zoom.

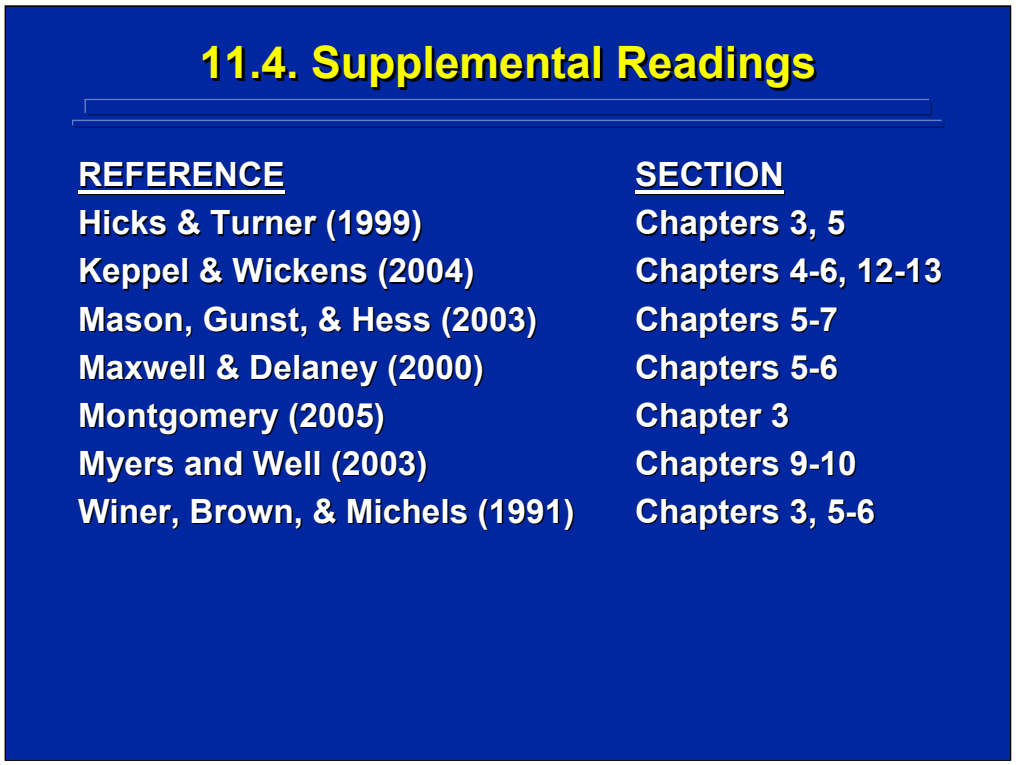
## 11.3. Summary

- Isolating ANOVA Main Effects and Interactions
  - Overall Test of Significance
  - Post Hoc Analyses
- Unplanned Paired Comparisons
  - Inflated $\alpha$ Error
  - Variety of Alternatives
  - Choice of Alternatives
- Interaction Analysis
  - Graphical Procedure
  - Analytical Procedures
  - Evaluation Process

By way of summary, this topic covered a variety of techniques that can be used to isolate main effects and interactions that are statistically significant in the overall ANOVA. The overall test confirms that at least one pair of means is significantly different, but post hoc analyses are needed to determine exactly which mean differences are significant in a main effect that has more than two levels or in an interaction effect.

Paired comparisons of treatment means are the most often used post hoc analysis of significant main effects and interactions. A variety of paired comparison procedures are available depending on the strategy chosen to control for inflated Type I error that occurs when multiple contrasts are performed on the same set of data.

Isolating the differential effect of an interaction involves both graphing and analytical procedures. Simple effects tests, trend analyses, and paired comparisons are appropriate analytical procedures. But, the primary analysis involves paired comparisons of the unconfounded contrasts of interaction treatments. The experimenter should always take care to support graphical representations of interactions with analytical procedures. Once the locus of the interaction is determined analytically, it should be interpreted in connection with any significant main effects.




Appropriate chapters in common experimental design textbooks used by human factors researchers are listed on this slide. Maxwell and Delaney (2000) provide a detailed discussion of various multiple comparison procedures and trend analysis in Chapters 5 and 6, respectively. The chapters in Keppel and Wickens (2004) and Winer et al. (1991) provide detailed discussions of linear comparisons, simple-effects tests, trend analysis, and alternative tests for paired comparisons.

# Topic 12. Within-Subjects ANOVA Designs

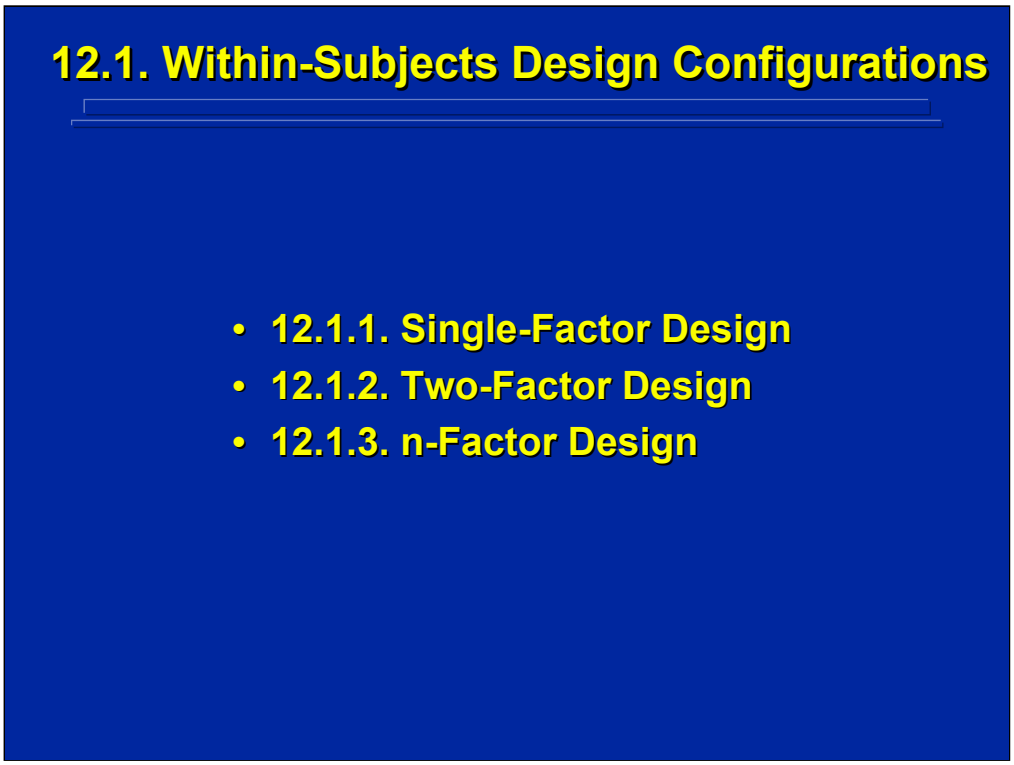
12.1. Within-Subjects Design Configurations
- 12.1.1. Single-Factor Design
- 12.1.2. Two-Factor Design
- 12.1.3. n-Factor Design

12.2. Homogeneity of Covariance
12.3. Balancing Order of Treatments
- 12.3.1. Balancing Alternatives
- 12.3.2. Balanced Latin Square
- 12.3.3. Testing Order Effects

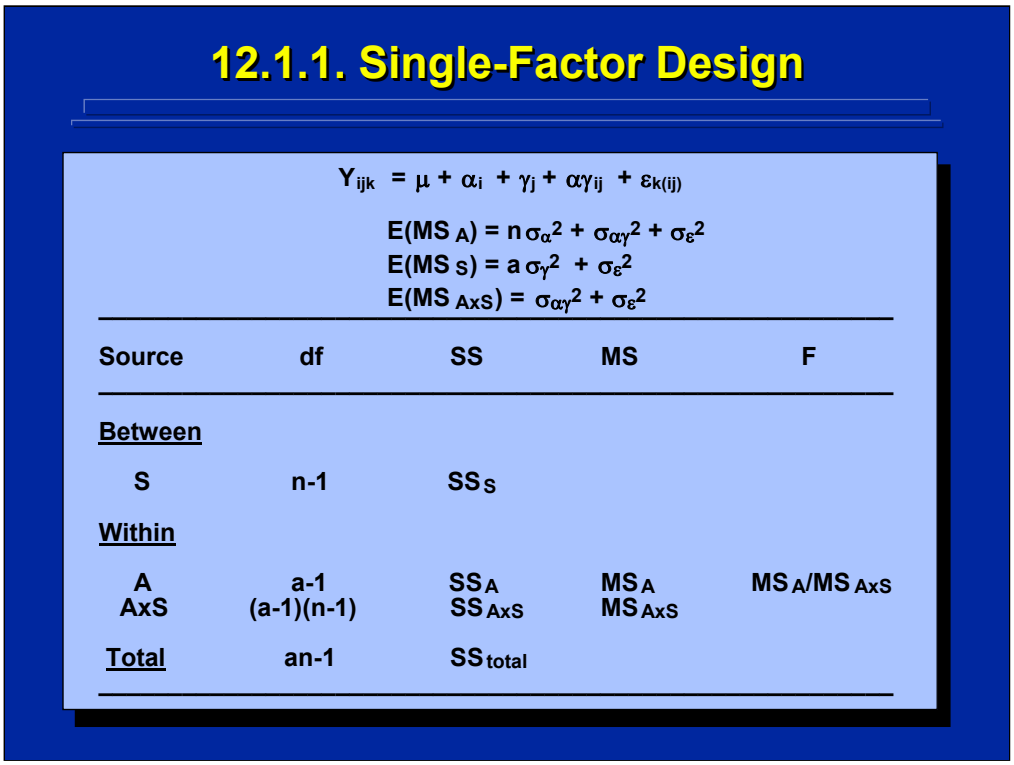
12.4. Differential Transfer
12.5. Within-Subjects Design Advantages
12.6. Summary
12.7. Supplemental Readings

This topic covers within-subjects ANOVA designs in which each subject receives every treatment condition in the human factors experiment. Basic configurations of one- and two-factor designs as well as generalizations to n-factor repeated measures designs are covered. The advantages of within-subjects designs are summarized as compared to between-subjects designs.

Additional considerations in using repeated measures are discussed including the homogeneity of covariance assumption, balancing techniques for controlling possible confounding effects of treatment orders, and the effect of differential transfer. References to supplemental readings on these issues as well as details on the design and analysis of within-subjects designs are provided in the major experimental design texts appropriate for human factors research.

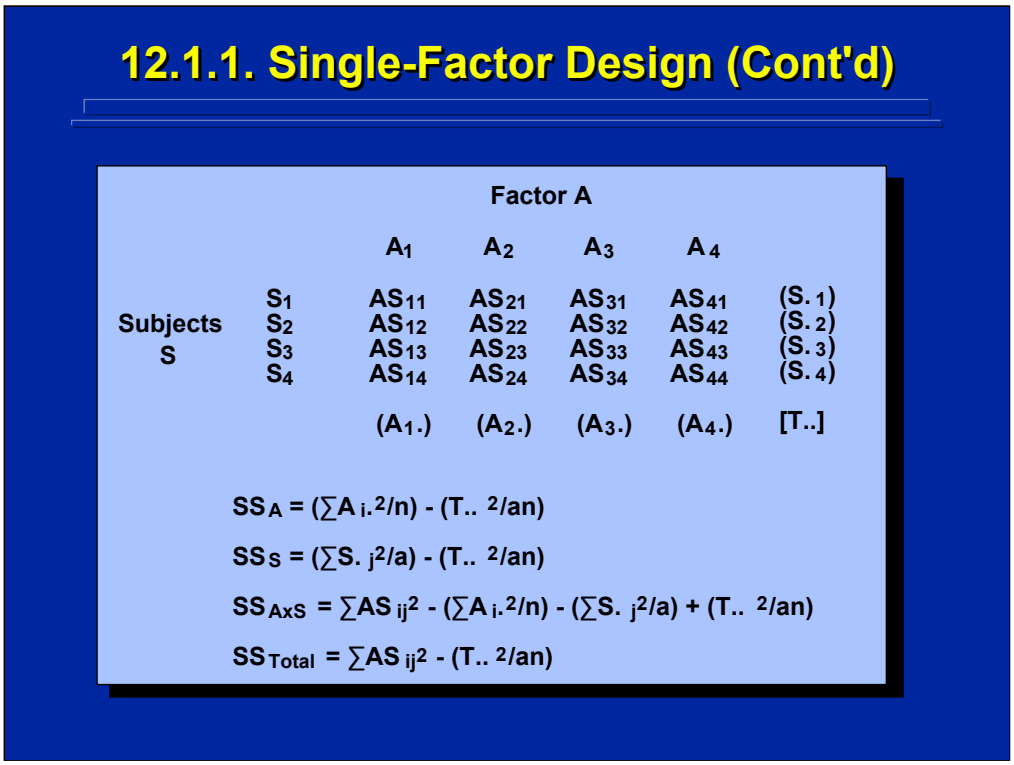



Within-subject design and analysis configurations are presented in this subsection using the simplified notation as well as the general rules, procedures, and algorithms for generating ANOVA designs as discussed for both one- and two-factor, between-subjects designs. Computational examples are provided for both one- and two-way, within-subjects designs. The SAS analyses for these examples are presented in Slater and Williges (2006) appendix. Based on the discussion of basic designs, generalizations are summarized for any n-factor within-subjects design.



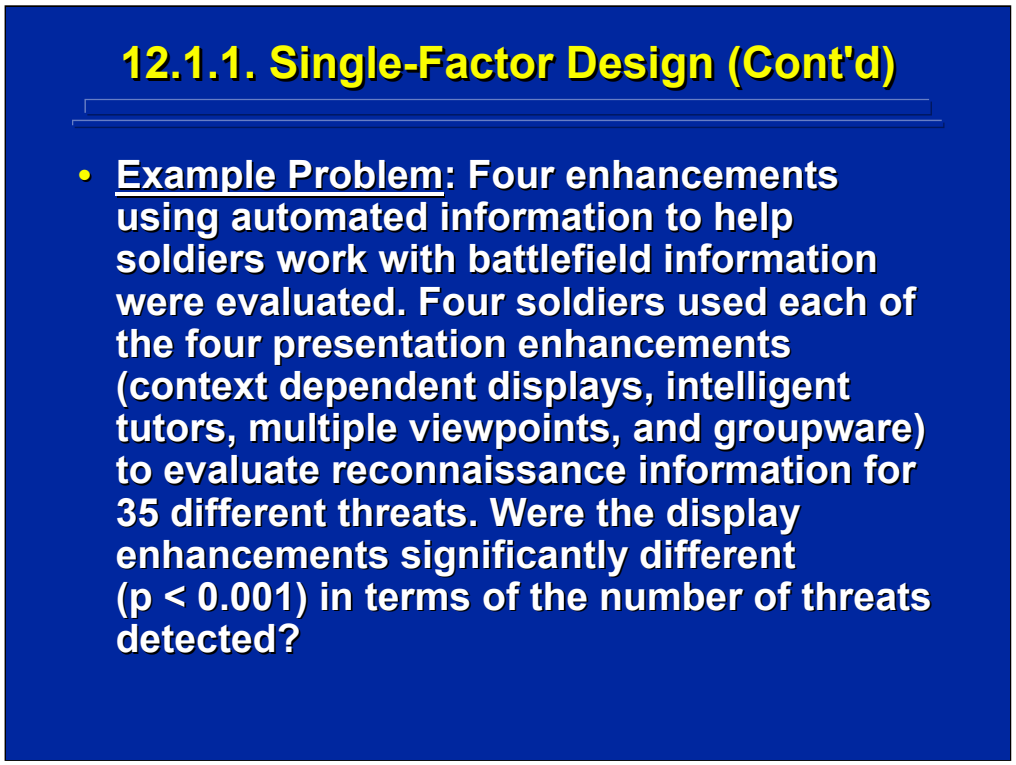

The fundamental within-subjects design involves only one factor in which each subject receives every level of that factor. The statistical model and expected mean squares are presented on the top portion of this slide. Note that subjects, S, are crossed with Factor A resulting in an AxS interaction.

The general form of the ANOVA Summary Table is presented in the lower portion of this slide. Based on the E(MS) for this design, the error term for testing the Factor A main effect is $MS_{AxS}$. There is no legitimate (unbiased) error term for testing the differences among subjects. Thus, the variability due to subjects is merely removed from the error term as a means of making the design more sensitive for testing Factor A. The df and SS for subjects are normally presented in the ANOVA Summary Table for completeness even though they are not used in any F-test.



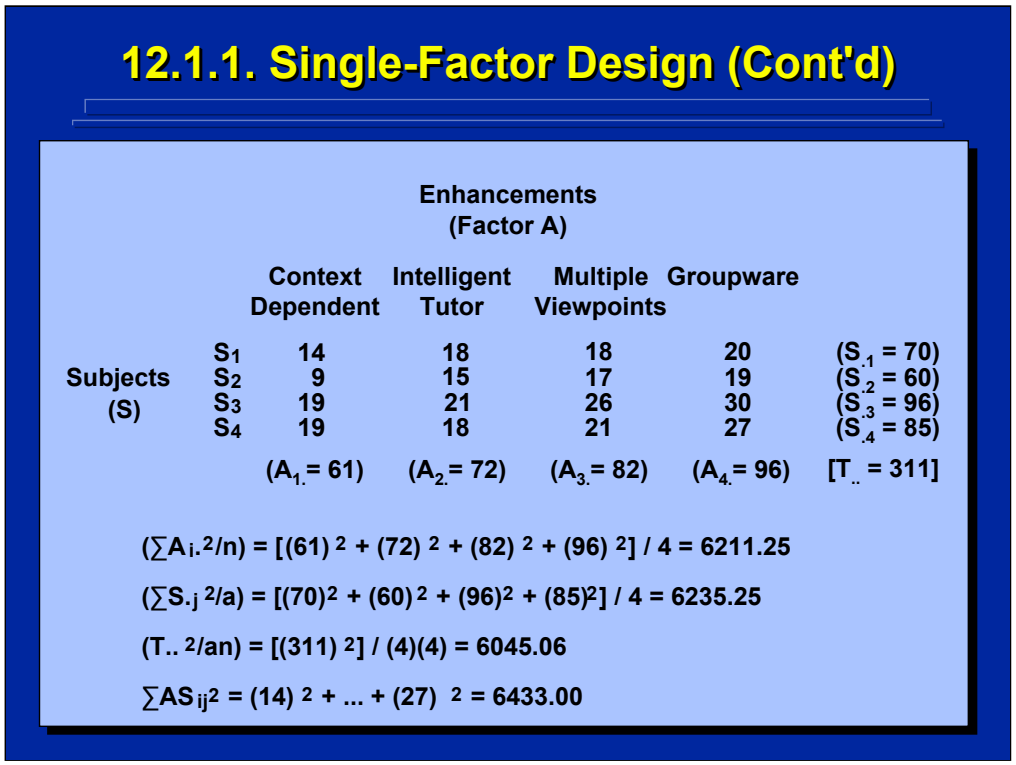

This slide shows the general layout of the data set for a one-way, within-subjects ANOVA design. Note that there are four different subjects shown in this layout, and each subject receives all four levels of Factor A.

The SS computational formulae for this design are shown on the bottom portion of this slide in the simplified dot notation. These formulae can be generated using the algorithm for SS formulae. Notice that the various formulae are composed of various combinations of the four computational components listed in the data set.



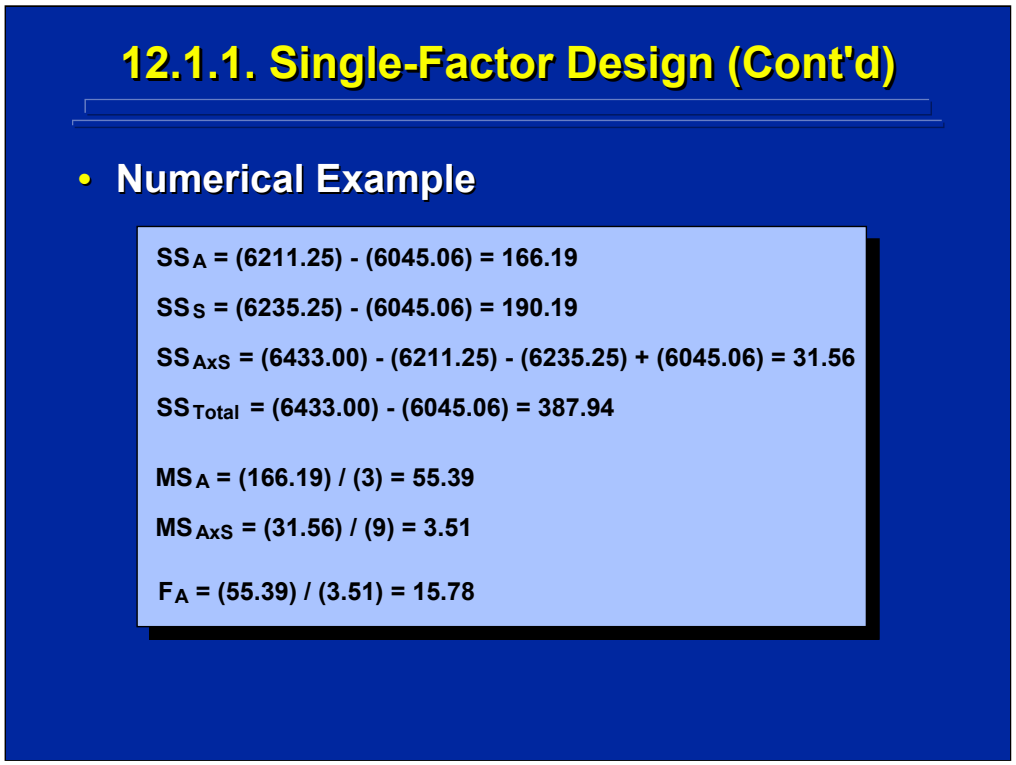

(Click in this red rectangle to see SAS calculations for this example.)

This slide presents an example problem that illustrates a one-way, within-subjects design. Since each soldier used each of the four automated information enhancements, this a repeated measures design. The Slater and Williges (2006) appendix provides the SAS analysis of this problem.




(Click in this red rectangle to see SAS calculations for this example.)

The data set for the example problem is shown on this slide. Both real-world descriptors and simplified notation designations are listed on the slide. Calculations of the four computational components of the SS are shown on the bottom portion of this slide.




(Click in this red rectangle to see SAS calculations for this example.)

This slide summarizes the calculations of the SS, MS, and F ratio in the example problem. These calculations are based on the data set and component values shown on the previous slide.

## 12.1.1. Single-Factor Design (Cont'd)

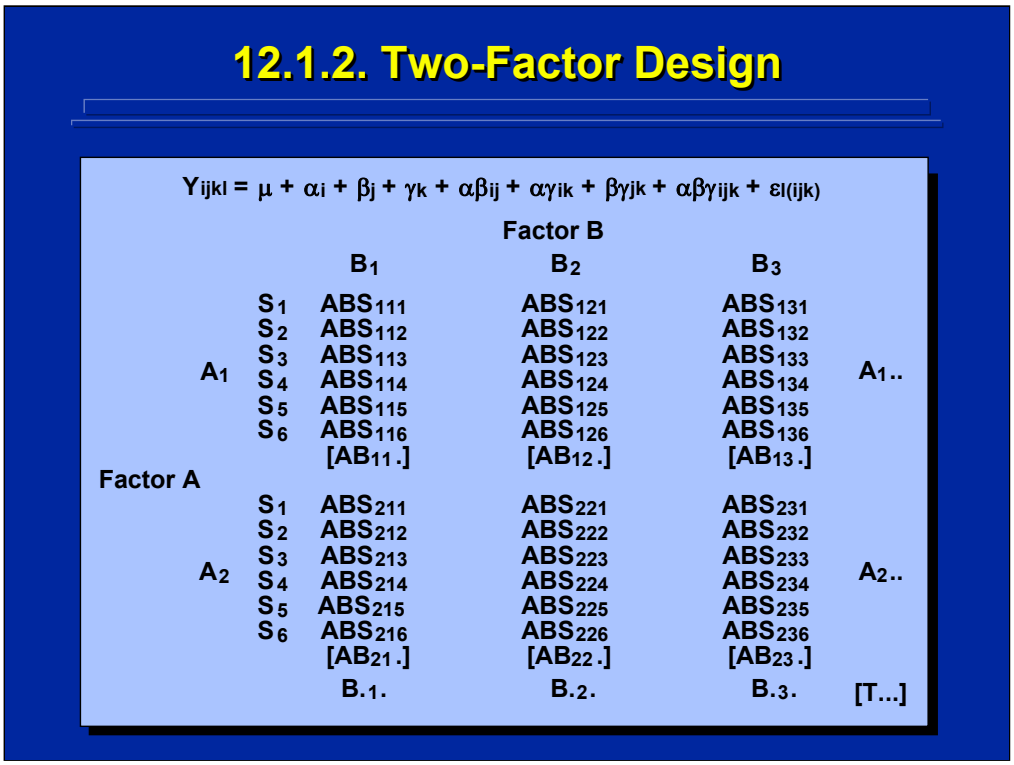
- **ANOVA Summary Table**

| Source | df | SS | MS | F |
|---|---|---|---|---|
| **Between** | | | | |
| Subjects (S) | 3 | 190.19 | | |
| **Within** | | | | |
| Enhancements (E) | 3 | 166.19 | 55.39 | 15. 78* |
| ExS | 9 | 31.56 | 3.51 | |
| **Total** | 15 | 387.94 | | |

*p < 0.001

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the ANOVA Summary Table of the one-factor, within-subjects design. Note that in the Summary Table for the example problem the Enhancement factor is stated as in the example problem rather than in the simplified notation. The total df are equal to 1 less than the 16 observations in the entire experiment, and $SS_{Total}$ equals the sum of all the SS components in the within-subjects design. Note that Enhancements is significant ($p < 0.001$) when compared to the tabled value, $F_{(3, 9)} = 13.90$). This means that at least one of the automated information enhancements is different from the others at the 0.001 level of significance. Subsequent, post-hoc, paired comparisons of the four types of information enhancements are needed to determine exactly which pairs are significantly different.




The statistical model listed at the top of this slide shows that subjects are crossed with both Factor A and Factor B to form a within-subjects design. This slide also shows the general layout of a two-factor, within-subjects design data set that is specified in the simplified notation. Sample size is six for the design shown on the slide, and the same six subjects are listed for levels $A_1$ and $A_2$ to designate a within-subjects design layout. In fact, each of these six subjects experiences all six treatment combinations in the 2x3 factorial design. Consequently, the sum for Subjects, $S_{..k}$, is summed over the six observations of each subject in the design.

## 12.1.2. Two-Factor Design (Cont'd)

- **SS Computational Formulae**

$$SS_S = (\sum S..{}_k{}^2/ab) - (T...{}^2/abn)$$

$$SS_A = (\sum A_i..{}^2/bn) - (T...{}^2/abn)$$

$$SS_{AxS} = (\sum AS_{i.k}{}^2/b) - (\sum A_i..{}^2/bn) - (\sum S..{}_k{}^2/ab) + (T...{}^2/abn)$$

$$SS_B = (\sum B._j.{}^2/an) - (T...{}^2/abn)$$

$$SS_{BxS} = (\sum BS._{jk}{}^2/a) - (\sum B._j.{}^2/an) - (\sum S..{}_k{}^2/ab) + (T...{}^2/abn)$$

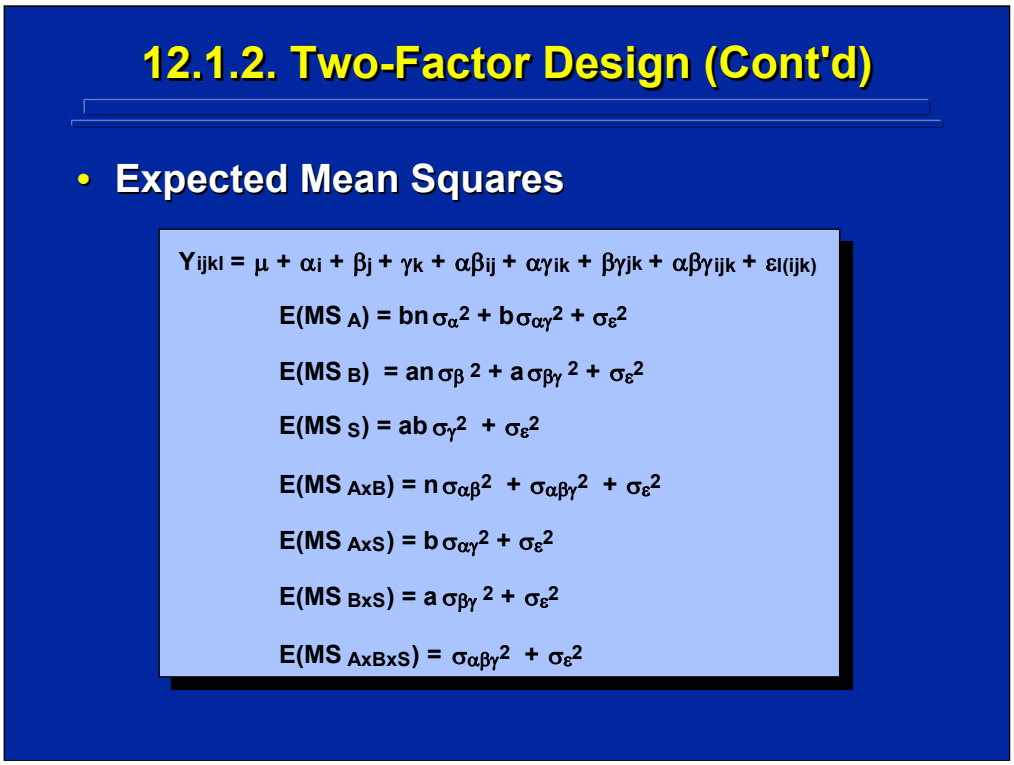
$$SS_{AxB} = (\sum AB_{ij}.{}^2/n) - (\sum A_i..{}^2/bn) - (\sum B._j.{}^2/an) + (T...{}^2/abn)$$

$$SS_{AxBxS} = \sum ABS_{ijk}{}^2 - (\sum AB_{ij}.{}^2/n) - (\sum AS_{i.k}{}^2/b) - (\sum BS._{jk}{}^2/a) + (\sum A_i..{}^2/bn) + (\sum B._j.{}^2/an) + (\sum S..{}_k{}^2/ab) - (T...{}^2/abn)$$

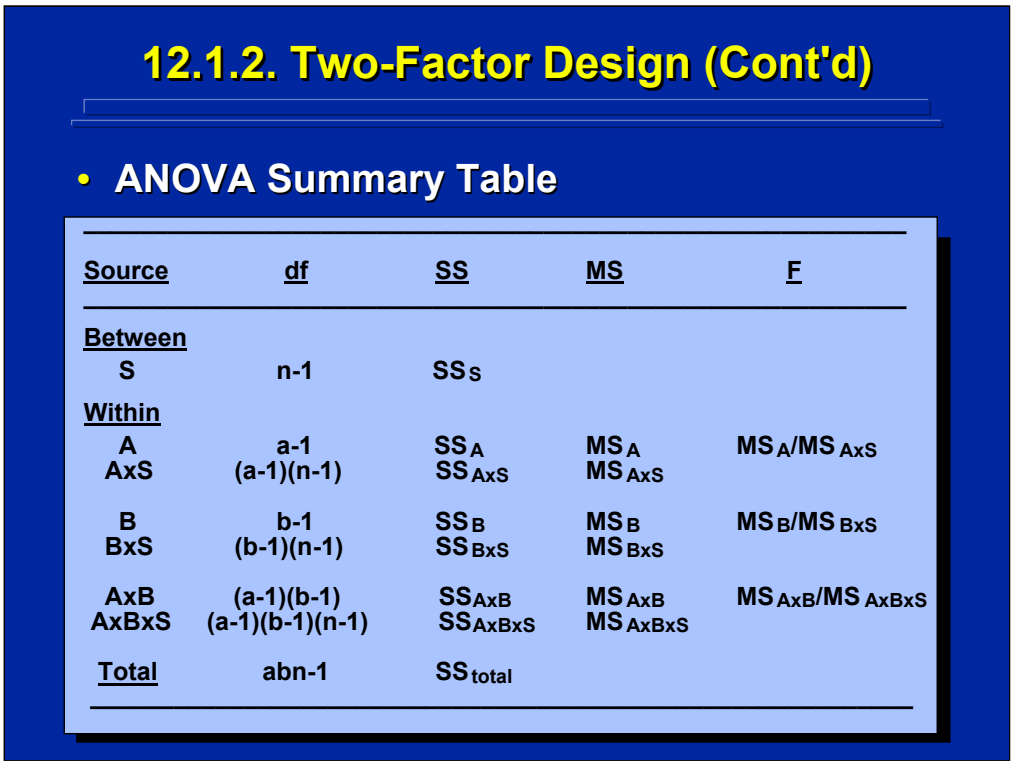
$$SS_{Total} = \sum ABS_{ijk}{}^2 - (T...{}^2/abn)$$

The complete SS computational formulae for the two-way, within-factor ANOVA design are listed on this slide. These formulae can be determined by using the SS algorithm.

Notice that there are eight different component scores that make up these SS formulae. The $S_{..k}$, the $AS_{i.k}$ and the $BS_{.jk}$ values are not shown on the previous data layout slide and need to be calculated in addition to the four values listed on the previous slide in order to calculate the eight components used in the SS calculations for this design.




The expected mean squares for a two-factor, within-subjects design are listed on this slide as determined by the E(MS) algorithm. Notice that the A and B main effects and the AxB interaction are divided by their respective interaction with subjects, S, to form a legitimate F-ratio. Based on the E(MS) designation, there is no legitimate (unbiased) error term to test the subject effect.



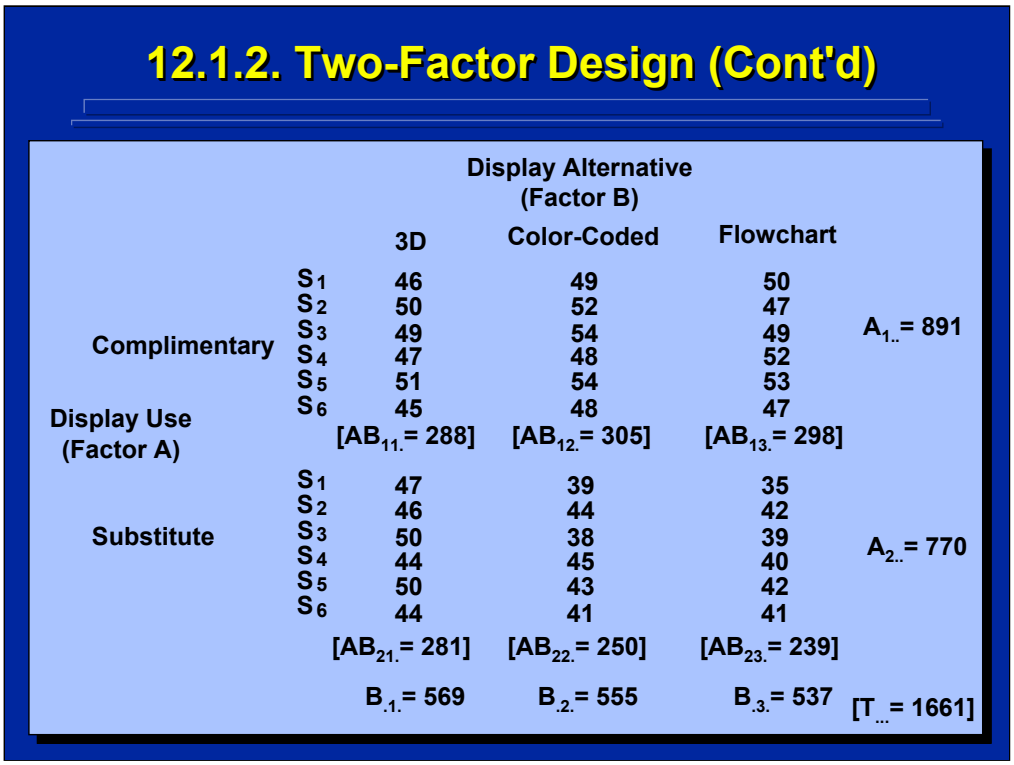

The complete ANOVA Summary Table for the two-factor, within-subjects design is shown on this slide. Only Factor A, Factor B, and the AxB interaction can be tested. Each of these effects is grouped with its error term and listed as a within-subjects effect. The main effect of subjects, S, is listed as a between-subjects effect for completeness and as a way of checking for computational errors when totaling df and SS.

## 12.1.2. Two-Factor Design (Cont'd)

- **Example Problem: Three alternative visual displays (3 dimensional graphs, color coded diagrams, and flowcharts) were developed to augment intelligence information gathered over a 12-hour period. Six intelligence officers evaluated the information using each visual display either as redundant to or as a substitute for the standard written intelligence information. Are the information presentations significantly different ($p < 0.05$)?**

(Click in this red rectangle to see SAS calculations for this example.)

A 2x3 within-subjects design is described on this slide as an example problem. Note that each of the six intelligence officers experienced each of the six treatment combinations resulting from the factorial combination of the two levels of display use (i.e. complimentary or substitute) and three levels of alternative visual displays of written intelligence information (i.e. 3D graphs, color-coded diagrams, and flowcharts). This reference material summarizes the calculations for the experiment; whereas the Slater and Williges (2006) appendix provide the SAS analysis for this example problem.




(Click in this red rectangle to see SAS calculations for this example.)

Hypothetical data are listed on this slide showing both the real-world designations of the factors and levels in the example problem as well as various totals listed in the simplified notation. Each of the six data points in each cell of the data set layout is an $ABS_{ijk}$ entry that represents one of the 36 data points in the experiment. In addition to the totals listed on the slide, $S_{..k}$, $AS_{i.k}$, and $BS_{.jk}$ totals are also needed for SS calculations.

## 12.1.2. Two-Factor Design (Cont'd)

- **ANOVA Summary Table**

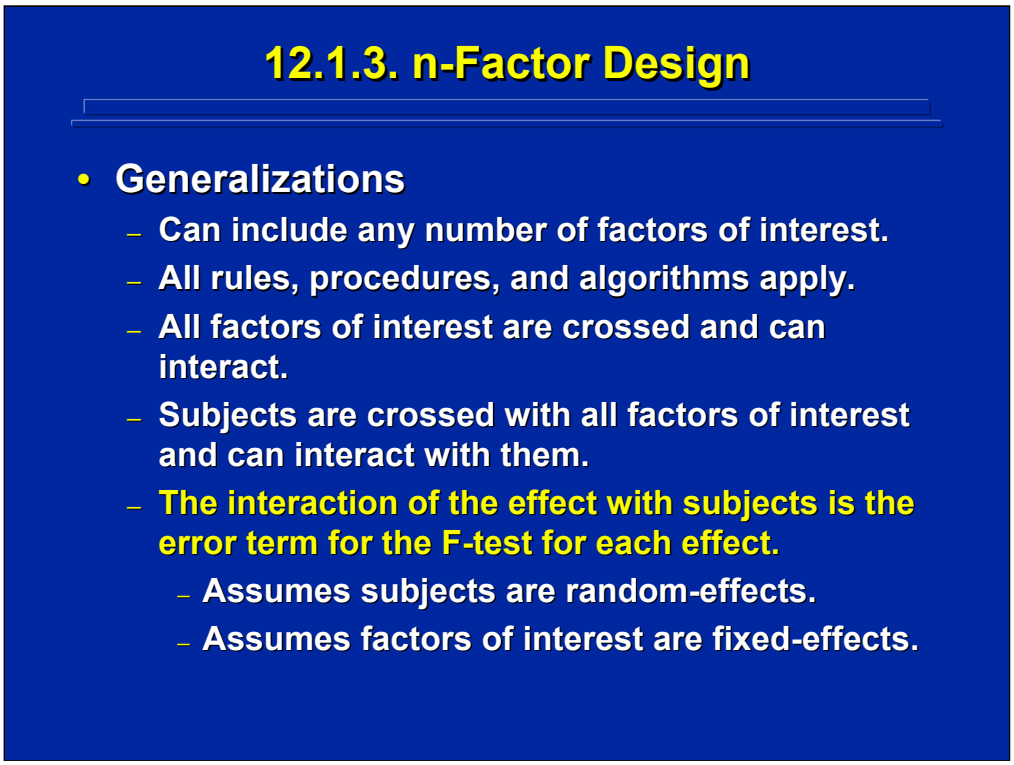
| Source | df | SS | MS | F |
|---|---|---|---|---|
| **Between** | | | | |
| Subjects (S) | 5 | 86.47 | 17.29 | |
| **Within** | | | | |
| Use (U) | 1 | 406.69 | 406.69 | 123.45*** |
| UxS | 5 | 16.47 | 3.29 | |
| Alternative (A) | 2 | 42.89 | 21.41 | 6.82* |
| AxS | 10 | 31.44 | 3.14 | |
| AxU | 2 | 139.56 | 69.78 | 9.09** |
| AxUxS | 10 | 76.78 | 7.68 | |
| Total | 35 | 800.31 | | |

*p < 0.05 **p < 0.01 ***p < 0.001

(Click in this red rectangle to see SAS calculations for this example.)

The complete ANOVA Summary Table for the two-factor, within-subjects design example problem is provided on this slide in standard format. Note that both the df and SS for all the effects in this design sum to the totals. In addition, the effects are grouped with their appropriate error terms for easy reference.

The main effect of the two display uses (U), the main effect of the three display alternatives (A), and the display use by display alternative (AxU) interaction are each significant at the 0.05 level when compared to the tabled F values. Since display use only has two levels, the experimenter can conclude that complimentary displays rather than substitution displays for written intelligence information resulted in significantly better intelligence. Further post-hoc analyses are needed to isolate differences among the three display alternatives and the AxU interaction.



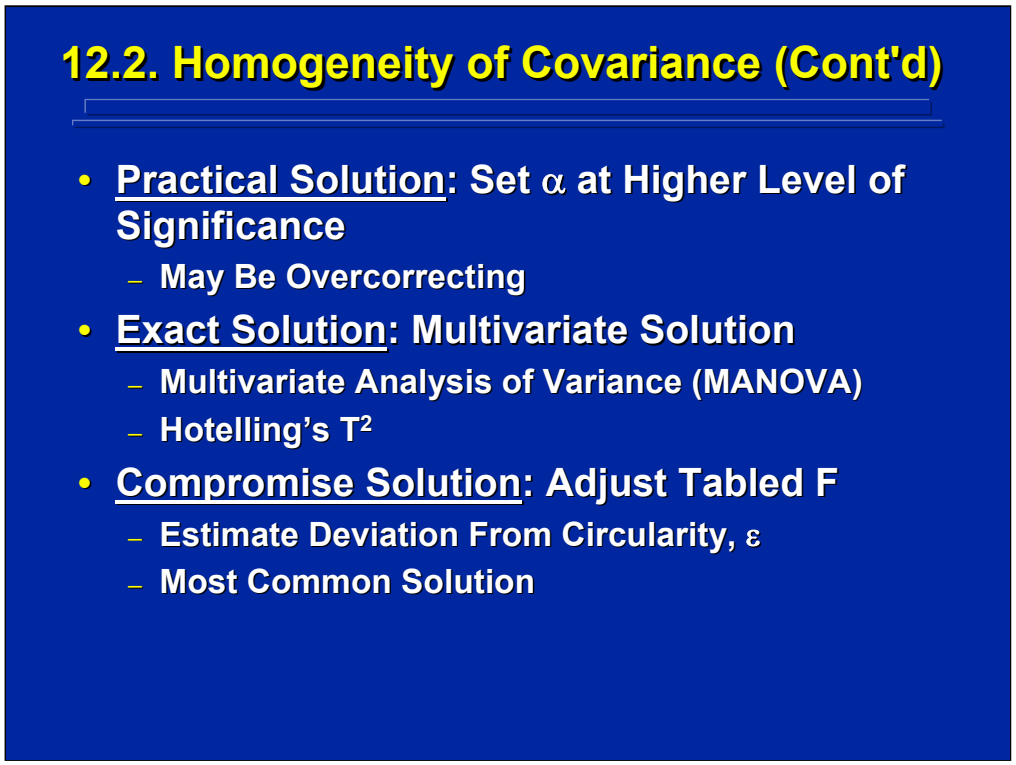

This slide provides generalizations for constructing and analyzing any n-factor, factorial within-subjects design with equal sample size. If the researcher assumes that the Subjects factor is the only random-effect variable and all factors of interest in the experiment are fixed-effects, then the error term for testing the effects of interest is simply the interaction of the effect with subjects.

## 12.2. Homogeneity of Covariance

- **Assumption of Homogeneity of Covariance: Covariance between pairs of treatment conditions is equal.**
  - **Heterogeneity Possible with More Than Two Levels**
  - **Calculate Intercorrelation Matrix**
  - **Positive Bias: $\alpha$ Error Increases**
- **Metrics from Population Variance-Covariance Matrix, $\Sigma_X$**
  - **Compound Symmetry: Equal Variances and Covariances + Circularity**
  - **Circularity: Sum of any two treatment variances minus their covariances is a constant.**
  - **Sphericity: Normalized Orthogonal Transformation to Orthonormal Variance-Covariance Matrix, $\Sigma_Y$**
  - **Departure from Circularity (Box, 1954): $\varepsilon$**

Since repeated observations are made on each subject in within-subjects designs, covariance exists among treatment levels. Within-subjects designs assume homogeneity of variance of within-treatment conditions as well as homogeneity of covariance among treatments in order for the observed F-ratio to be distributed according to the F sampling distribution. If more than two levels of repeated measures exist, there is the possibility of heterogeneity of covariance in within-subjects designs. Covariance among repeated treatments can be specified by unequal correlation between them. Violation of the assumption of homogeneity of covariance results in a positive bias in the F-test which yields an increase in Type I error.

Winer et al. (1991, pp. 237-282 and pp. 509-526) provide an excellent mathematical discussion of the homogeneity of covariance assumption and corrections for heterogeneity of covariance when it exists. Various terms based on the population variance-covariance matrix as shown on the bottom of this slide are used to assess heterogeneity of covariance. Compound symmetry expresses homogeneity of variance and covariance, but this criterion is usually relaxed to just estimates of circularity. Consequently, adjusted F-tabled values corrected for deviations from circularity, $\varepsilon$, based on sample data are often used when a violation of the homogeneity of covariance assumption is of concern to the experimenter.




There are three alternatives to consider when heterogeneity of covariance exits in the within-subjects data set. First, the experimenter can choose to test the within-subjects design at a higher level of significance (0.01 instead of 0.05) to guard against inflated $\alpha$ error when using the F sampling distribution. But, this approach could be too stringent if the real intent is to test at a lower $\alpha$ level.

Second an exact solution to the ANOVA that includes the degree of covariance among treatment conditions can be calculated using multivariate analysis of variance (MANOVA). In Chapters 13 and 14, Maxwell and Delaney (2000) discuss the use of MANOVA as a multivariate procedure to provide an exact solution for repeated measures designs. Winer et al. (1991, pp. 278-281) discusses the use of a multivariate Hoetelling's $T^2$ as an exact solution for one-way designs.

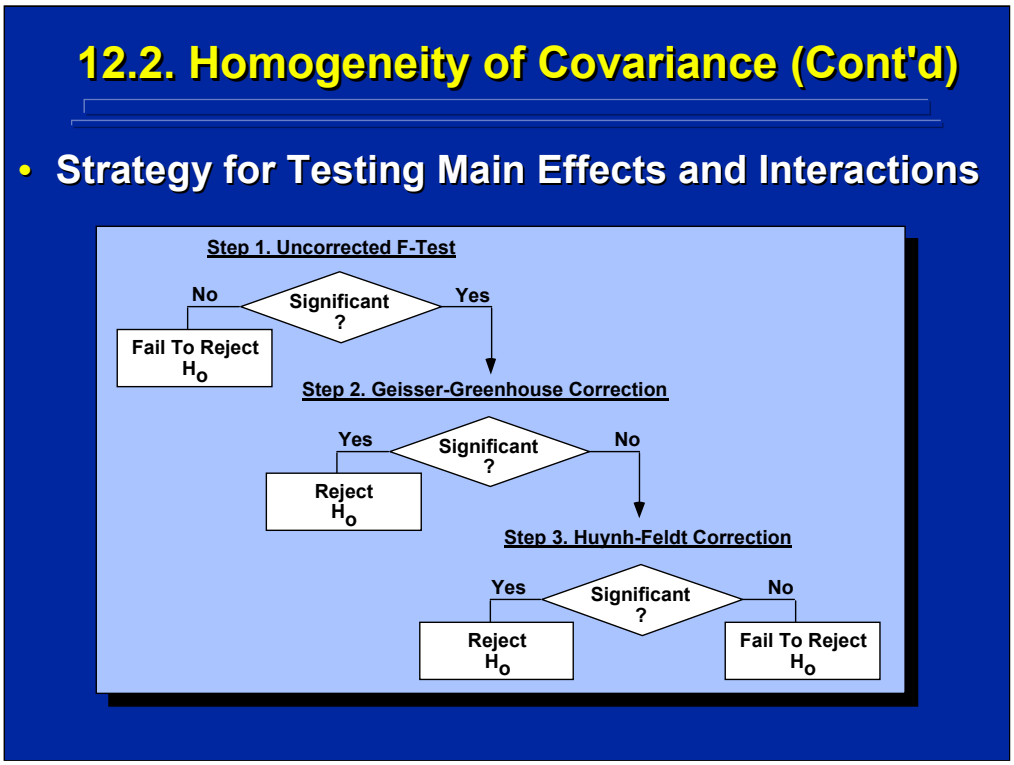
The third alternative is a compromise solution that uses the standard univariate ANOVA computations but adjusts the tabled F value based on the lack of circularity. This approach is summarized in this reference material because it is the approach commonly used in human factors research.

## 12.2. Homogeneity of Covariance (Cont'd)

- Range of Deviation From Circularity, ε
  - 1/(k-1) = ε ≤ 1 where,
    - k = Number of Treatment Levels
    - 1 = Circularity Assumption Met
- Geisser-Greenhouse (1958) Maximum Correction
  - Adjusted F Table
    - F(1, n-1)
- Box Correction for Repeated Measures (Box, 1954)
  - Adjusted F Table
    - F[(k-1)ε, (n-1)(k-1)ε]
  - Imhof (1962) Table for Small Sample Size (n<9)
- Huynh and Feldt (1976) Estimate of ε from Sample Data
  - Statistical Packages for Correction Computations
- No Correction for Unplanned Comparisons

Adjustments to the standard F table are based on the amount of deviation from circularity, ε, that exists in the data set. The formula at the top of this slide provides the possible range of deviation. Geisser and Greenhouse (1958) consider only the maximum deviation from circularity for an adjustment to the F table. Their approach always results in an overcorrection for repeated measures unless heterogeneity of covariance is the maximum.

Box (1954) provided an adjusted F tabled value based on the value of ε as shown in the middle of this slide. Alternatively, the Imhof (1962) Table is available in the appendix of Winer et al. (1991) as Table D.18 that can be used for the Box correction when sample size is less than 9. To use the Box correction, the experimenter needs to estimate ε from the sample data. The Huynh and Feldt (1976) value is the most commonly used estimate of ε and is a correction based on the Collier, Baker, Manville, and Hayes (1967) formula to estimate ε based on sample data. Note that this computation becomes complex as shown in Winer et al. (1991, pp. 253), and statistical packages are usually used for this calculation. This correction is used for testing main effects and interactions in ANOVA, but subsequent unplanned, post hoc comparisons are made without correction.




If the experimenter suspects marked deviations from circularity, several choices are available for heterogeneity of covariance. The maximum correction, Geisser-Greenhouse, can be adopted as a conservative approach. Alternatively, the Huynh-Feldt correction can be used as a more exact solution based on sample data estimates of $\varepsilon$.

This slide diagrams a general strategy using various corrections that the experimenter might use in correcting within-subjects design ANOVAs that violate the homogeneity of covariance assumption similar to the approach described by Myers and Wells (2003, p. 359). First an uncorrected F-test is conducted which may have a positive bias. If the result is not significant, analysis stops. If the result is significant, then it is retested using the Geisser-Greenhouse maximum protection for deviation from circularity. If the test is significant, further analysis stops and the experimenter rejects the null hypothesis. If the Geisser-Greenhouse correction is not significant, then the experimenter uses the Huynh-Feldt correction based on sample estimate of circularity deviations.

## 12.2. Homogeneity of Covariance (Cont'd)

- One-Factor, Within-Subjects Design Example

| Source | df | SS | MS | F | p | G-G p | H-F p |
|---|---|---|---|---|---|---|---|
| Between | | | | | | | |
| Subjects (S) | 3 | 190.19 | | | | | |
| Within | | | | | | | |
| Enhancements (E) | 3 | 166.19 | 55.39 | 15.80 | 0.0006 | 0.0053 | 0.0006 |
| ExS | 9 | 31.56 | 3.51 | | | | |
| Total | 15 | 387.94 | | | | | |

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the Geisser-Greenhouse and Huynh-Feldt corrections for the one-factor, within-subjects design, example problem as calculated by SAS and described in the Slater and Williges (2006) appendix. Following the F-test strategy diagrammed in the previous slide, the experimenter would calculate both corrections and conclude the main effect of Enhancements is significant at the 0.001 level of significance based on the Huynh-Feldt correction. Note that the Geisser-Greenhouse maximum correction (p = 0.0053) is more severe than the Huynh-Feldt correction (p = 0.0006).

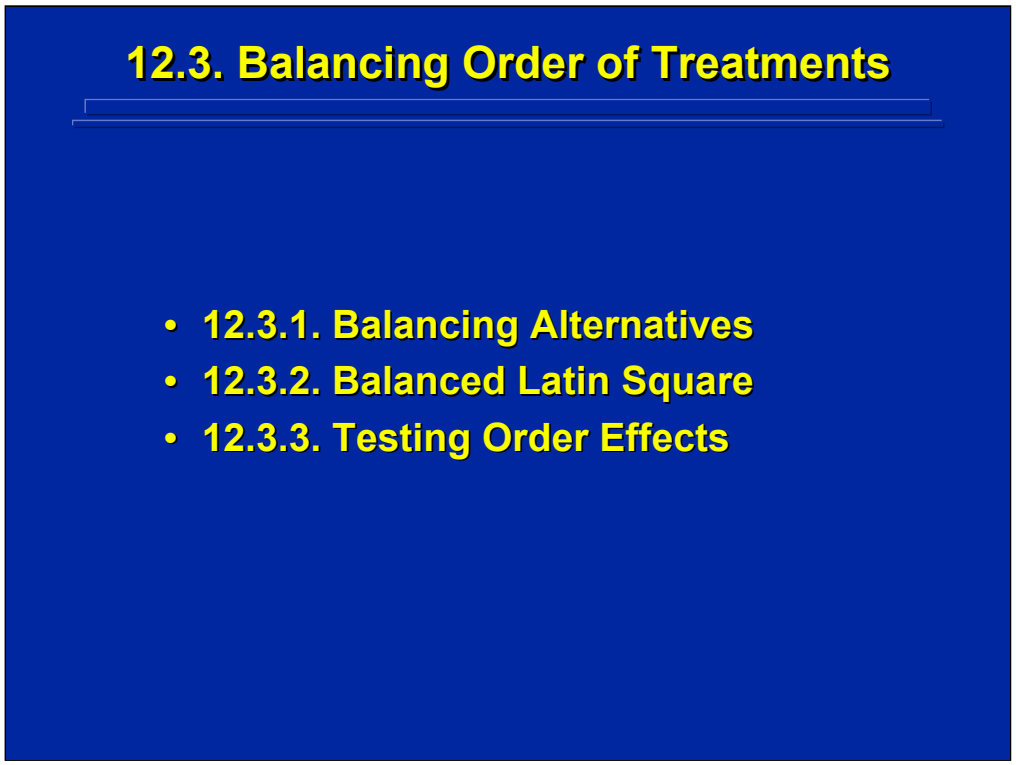
## 12.2. Homogeneity of Covariance (Cont'd)

- **Two-Factor, Within-Subjects Design Example**

| Source | df | SS | MS | F | p | G-G p | H-F p |
|---|---|---|---|---|---|---|---|
| **Between** | | | | | | | |
| Subjects (S) | 5 | 86.47 | 17.29 | | | | |
| **Within** | | | | | | | |
| Use (U) | 1 | 406.69 | 406.69 | 123.45 | 0.0001 | 0.0001 | 0.0001 |
| UxS | 5 | 16.47 | 3.29 | | | | |
| Alternative (A) | 2 | 42.89 | 21.41 | 6.82 | 0.0135 | 0.0303 | 0.0202 |
| AxS | 10 | 31.44 | 3.14 | | | | |
| AxU | 2 | 139.56 | 69.78 | 9.09 | 0.0056 | 0.0062 | 0.0056 |
| AxUxS | 10 | 76.78 | 7.68 | | | | |
| Total | 35 | 800.31 | | | | | |

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows both Geisser-Greenhouse and Huynh-Feldt corrections for heterogeneity of covariance for the two-factor, within-subjects design example problem. These corrected p-values were calculated by SAS as described in the Slater and Williges (2006) appendix. Note that the correction is made for the F-test on both main effects and the AxU interaction. As shown in the previous one-way example, the Geisser-Greenhouse correction is the maximum correction for heterogeneity of covariance and results in a lower significance level than the Huynh-Feldt correction. Note that factor Use does not have a correction for sphericity because that factor has only two levels. In any event, all three F-tests are significant at the 0.05 level of significance.




A major procedural component of any within-subjects ANOVA design is to choose a technique for balancing the presentation order across subjects for the within-subjects treatments so that practice order is not confounded with the treatment effects. Cotton (1998) discusses the importance of balancing treatment orders to control carryover effects and describes various design alternatives for balancing and testing carryover effects in repeated measures experiments. Various procedures for balancing presentation orders in factorial ANOVA designs are presented in this subsection.

## 12.3. Balancing Order of Treatments (Cont'd)

- **Totally Confounded Presentation Order of "t" Levels of Treatments**
  - Every Subject Receives the Same Order of Treatments
- **Example: One-Factor, Within-Subjects Design with Three Levels**

| | | Subjects | | |
|---|---|---|---|---|
| | | $S_1$ | $S_2$ | $S_3$ |
| Presentation Order | 1 | $A_1$ | $A_1$ | $A_1$ |
| | 2 | $A_2$ | $A_2$ | $A_2$ |
| | 3 | $A_3$ | $A_3$ | $A_3$ |

To illustrate the importance of balancing, this slide shows an ordering in which presentation order is totally confounded with the three levels of treatments in a one-way, within-subjects design. Note that level $A_1$ is always presented first followed by $A_2$ then $A_3$ for each of the three subjects in the experiment. Practice and treatments are totally confounded. Consequently, the experimenter cannot determine if any significant differences in Fact A are due to treatment or practice effects. Obviously, this type of balancing across subjects should always be avoided, unless the within-subjects factor of interest is practice (i.e., three practice trials) and balancing of presentation order is not necessary.

## 12.3.1. Balancing Alternatives

- **Completely Counterbalanced Presentation Order of "t" Levels of Treatments**
  - Possible Presentation Orders = t!
  - Requires a Minimum of t! Subjects
- **Example: One-Factor, Within-Subjects Design with Three-Levels**
  - t! = 3x2x1 = 6 Subjects

| | | Subjects | | | | | |
|---|---|---|---|---|---|---|---|
| | | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ |
| Presentation Order | 1 | $A_1$ | $A_3$ | $A_2$ | $A_1$ | $A_2$ | $A_3$ |
| | 2 | $A_2$ | $A_1$ | $A_3$ | $A_3$ | $A_1$ | $A_2$ |
| | 3 | $A_3$ | $A_2$ | $A_1$ | $A_2$ | $A_3$ | $A_1$ |

The best balancing alternative is to completely counterbalance all possible presentation orders of treatment levels across subjects. There are t! (t factorial) ways of presenting "t" treatment levels. So, a completely counterbalanced within-subjects experimental design requires a minimum of t! subjects.

For example, the three-level, within-subjects design shown on this slide has three factorial (3!) or six possible orders of the three treatment levels requiring a minimum of six subjects for complete counterbalancing. Consequently, the experimenter should choose multiples of six subjects when determining sample size for this experiment.

## 12.3.1. Balancing Alternatives (Cont'd)

- **Random Assignment of "t" Levels of Treatments across Subjects**
  - Total Treatment Orders too Large for Available Subjects
  - Use Random Number Table for Assignment
- **Example: 3x4x5 Within-Subjects Design**
  - $A_iB_jC_k$ Treatment Combinations = t = 60
  - Total Counterbalancing = 60! Treatment Orders
  - Total Counterbalancing Not Feasible
  - Choose Appropriate Sample Size (n)
  - Random Order of "t" Treatments Assigned to Each of "n" Subjects

In most human factors experiments the resulting number of treatment conditions is too large to allow complete counterbalancing. When the number of treatment orders is extremely large, the experimenter must resort to random assignment of treatment orders to subjects in the within-subjects design as a means of controlling order effects. In this situation, a random number table is used to determine the treatment order for each subject.

Consider the extremely large 3x4x5 within-subjects design shown on this slide. There are 60 different treatments in this factorial design that each subject receives in the experiment. The number of possible treatment orders is 60!. Obviously, counterbalancing is not feasible across subject, and random assignment of treatment order across the appropriate sample size of subjects is the only feasible approach. Most human factors experiments, however, result in a total number of treatment conditions that can be partially balanced as a compromise between the complete counterbalancing and random assignment alternatives.

## 12.3.1. Balancing Alternatives (Cont'd)

- **Partially Counterbalanced Design**
  - Compromise Approach
- **Example: Four-Level, Within-Subjects Design**
  - t! = 4! = 24 Subjects for Complete Counterbalancing
  - Partial Counterbalancing with Four Subjects

| | | Subjects | | | |
|---|---|---|---|---|---|
| | | $S_1$ | $S_2$ | $S_3$ | $S_4$ |
| | 1 | $A_1$ | $A_2$ | $A_3$ | $A_4$ |
| Presentation | 2 | $A_2$ | $A_3$ | $A_4$ | $A_1$ |
| Order | 3 | $A_3$ | $A_4$ | $A_1$ | $A_2$ |
| | 4 | $A_4$ | $A_1$ | $A_2$ | $A_3$ |

This slide shows one example of a partially counterbalanced design. In this four-level, within-subjects design, there are 4! or a total of 24 possible orders of the four treatment conditions. This requires a minimum of 24 subjects for complete counterbalancing where all possible orders of treatment presentations would occur. The partial counterbalancing shown on this slide only requires four subjects. Across the four subjects, note that each of the four levels of the within-subjects factor are presented once in each of the four presentation positions. Using multiples of four subjects to select the sample size would maintain the partial counterbalancing. Note, however, that the sequence of preceding and following treatments conditions is held constant and not balanced in this partially counterbalanced scheme (e.g., $A_2$ always follows $A_1$).

The partial counterbalancing shown on this slide is based on a Latin square design that is described in detail in Topic 18 in this reference material. Keppel and Wickens (2004, pp. 383-393) provide a general discussion of the design and analysis of Latin squares for balancing the carryover of order effects in within-subjects designs.

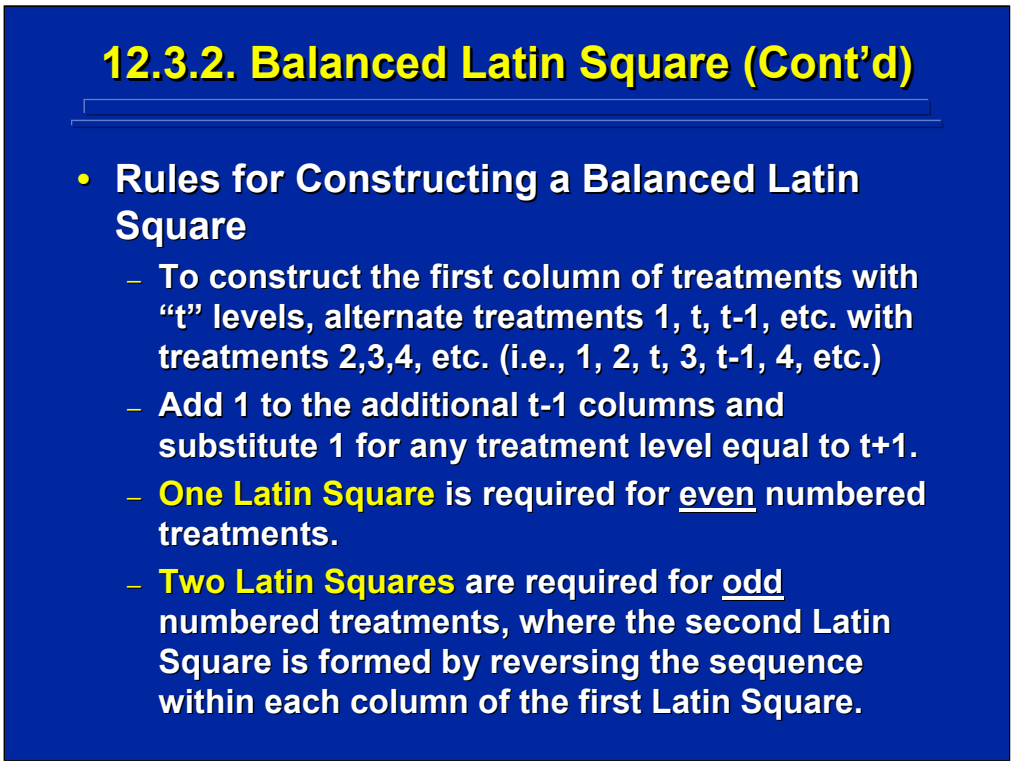
## 12.3.2. Balanced Latin Square

- **Balanced Latin Square Design Alternative for Ordering "t" Treatments across Subjects**
  - **Partial Counterbalancing of Sequence Effects**
  - **Requires "t" Subjects**
- **Example: Four-Level, Within-Subjects Design**

| | | Subjects | | | |
|---|---|---|---|---|---|
| | | $S_1$ | $S_2$ | $S_3$ | $S_4$ |
| | 1 | $A_1$ | $A_2$ | $A_3$ | $A_4$ |
| Presentation | 2 | $A_2$ | $A_3$ | $A_4$ | $A_1$ |
| Order | 3 | $A_4$ | $A_1$ | $A_2$ | $A_3$ |
| | 4 | $A_3$ | $A_4$ | $A_1$ | $A_2$ |

A special case of Latin squares called a balanced Latin square balances presentation order effects and some sequence effects. A balanced Latin square design is the most often used procedure in human factors research for balancing the order and sequence effects across the treatment effects in within-subjects designs. This procedure uses the same number of Subjects (S) as the Treatments (T) and Presentation Order (O) to construct the Latin square. The general format for presenting the balanced Latin square is to list S as the columns, O as the rows, and T as the entries within the Latin square.

This slide provides an example of a balanced Latin square scheme for a four-level, within-subjects design. Note that each treatment appears once in each presentation order, and each treatment precedes and follows the other treatments once across subjects. Such a partial counterbalancing scheme of order and some sequence effects allows presentation order to be independent of treatments in a within-subjects design. It also places restrictions on the choice of sample size (n) for the within-subjects design. Namely, the experimenter would choose multiples of four subjects in order to use a balanced Latin square in this four levels example. Although the number of subjects is the same as in the previous partial counterbalancing example, the balanced Latin square procedure provides more control over sequence order effects.




This slides lists the rules for constructing a balanced Latin square design consisting of "t" treatment levels based on procedures presented by Wiiliams (1949, 1950). Essentially, a balanced Latin square can be constructed by ordering the first column and then adding 1 to each treatment level in succeeding columns. The order of the first column is determined by alternating 1, t, t-1, etc. with treatment levels 2, 3, 4, etc. For example, if the number of treatment levels is four (i.e., t = 4) in the within-subjects design, the first column of the balanced Latin square is 1, 2, 4, 3. Note that only one Latin square is needed when there is an even number of treatment levels, and two Latin squares are needed when there is an odd number of treatment levels.

## 12.3.2. Balanced Latin Square (Cont'd)

- **Even Number of Treatments**
- **Example: 2x3 Within-Subjects Design**
  - Total of 6 $A_iB_j$ Treatment Combinations = t
  - Total Treatment Orders = t! = 6! = 720 Orders
  - Single 6x6 Balanced Latin Square Alternative
  - Columns = Subjects
  - Rows = Presentation Order

| First Column | Subjects | | | | | |
|---|---|---|---|---|---|---|
| | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ |
| 1 | 1 | 2 | 3 | 4 | 5 | 6 |
| 2 | 2 | 3 | 4 | 5 | 6 | 1 |
| t | 6 | 1 | 2 | 3 | 4 | 5 |
| 3 | 3 | 4 | 5 | 6 | 1 | 2 |
| t-1 | 5 | 6 | 1 | 2 | 3 | 4 |
| 4 | 4 | 5 | 6 | 1 | 2 | 3 |

This slide provides an example of using the rules for generating a balanced Latin square when the resulting number of treatments is an even number. In the 2x3 within-subjects design example, each subject receives a total of six treatments (i.e., t = 6) consisting of the factorial combination of $A_iB_j$ levels. There is a total of 720 (i.e., 6!) possible orders of these six combinations. So, complete counterbalancing is not feasible because it would require 720 subjects in the experiment. However, a balanced Latin square order of presentation of the resulting six treatment combinations across 6 subjects is a feasible alternative.

Application of the rules for generating this 6x6 balanced Latin square is shown on the bottom of this slide. Note that this balanced Latin square can be used to determine the order of the $A_iB_j$ treatment combinations in the two-factor, within-subjects design that each of the six subjects receives. The experimenter should use multiples of six subjects when choosing sample size for this example experiment in order to use a balanced Latin square for determining treatment presentation order in the within-subjects design example.

## 12.3.2. Balanced Latin Square (Cont'd)

- **Odd Number of Treatments**
- **Example: Seven-Level, Within-Subjects Design**
  - Total Treatment Orders = t! = 7! = 5,040 Orders
  - Two 7x7 Balanced Latin Squares Alternative
  - Minimum of 14 Subjects
  - Columns = Subjects
  - Rows = Presentation Order

| First Column | Subjects | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ | $S_7$ | $S_8$ | $S_9$ | $S_{10}$ | $S_{11}$ | $S_{12}$ | $S_{13}$ | $S_{14}$ |
| 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 5 | 6 | 7 | 1 | 2 | 3 | 4 |
| 2 | 2 | 3 | 4 | 5 | 6 | 7 | 1 | 4 | 5 | 6 | 7 | 1 | 2 | 3 |
| t | 7 | 1 | 2 | 3 | 4 | 5 | 6 | 6 | 7 | 1 | 2 | 3 | 4 | 5 |
| 3 | 3 | 4 | 5 | 6 | 7 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 1 | 2 |
| t-1 | 6 | 7 | 1 | 2 | 3 | 4 | 5 | 7 | 1 | 2 | 3 | 4 | 5 | 6 |
| 4 | 4 | 5 | 6 | 7 | 1 | 2 | 3 | 2 | 3 | 4 | 5 | 6 | 7 | 1 |
| t-2 | 5 | 6 | 7 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

This slide provides an example of using the rules for generating a balanced Latin square based on an odd number of treatment levels. In the seven-level, within-subjects design example, two 7x7 balanced Latin squares are generated in which the second balanced Latin square is the inverse of the first. The resulting balancing is such that each treatment level appears twice in each presentation order, and each treatment precedes and follows every other treatment twice across the 14 subjects participating in the experiment.

## 12.3.3. Testing Order Effects

- **Example: One-Way, Within-Subjects Design with Four Levels**
- **Balanced Latin Square**

| | | Subjects (S) | | | |
|---|---|---|---|---|---|
| | | $S_1$ | $S_2$ | $S_3$ | $S_4$ |
| Presentation Order (O) | $O_1$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ |
| | $O_2$ | $A_2$ | $A_3$ | $A_4$ | $A_1$ |
| | $O_3$ | $A_4$ | $A_1$ | $A_2$ | $A_3$ |
| | $O_4$ | $A_3$ | $A_4$ | $A_1$ | $A_2$ |

(Click in this red rectangle to see SAS calculations for this example.)

A balanced Latin square is used to keep possible order effects, if they exist, independent of the treatments effects of interest. The experimenter can conduct a subsequent ANOVA on the balanced Latin square to test for possible significant order effects if desired. To illustrate this procedure, the balanced Latin square shown on this slide is used for ordering treatments across 4 subjects in the one-way, within-subjects design example of battlefield information enhancement procedures that was presented at the beginning of this topic. The SAS procedures for conducting the ANOVA on a balanced Latin square is presented in the Slater and Williges (2006) appendix.

## 12.3.3. Testing Order Effects (Cont'd)

- **Assignment Data Matrix**

| | | Subjects (S) | | | | |
|---|---|---|---|---|---|---|
| | | $S_1$ | $S_2$ | $S_3$ | $S_4$ | |
| | $O_1$ | 14 | 15 | 26 | 27 | 82 |
| Presentation | $O_2$ | 18 | 17 | 30 | 19 | 84 |
| Order (O) | $O_3$ | 20 | 9 | 21 | 21 | 71 |
| | $O_4$ | 18 | 19 | 19 | 18 | 74 |
| | | 70 | 60 | 96 | 85 | [311] |

| Factor A Treatments (A) | | | | |
|---|---|---|---|---|
| $A_1$ | $A_2$ | $A_3$ | $A_4$ | |
| 61 | 72 | 82 | 96 | |

(Click in this red rectangle to see SAS calculations for this example.)

The data from the example of a one-way, within-subjects design in Section 11.1.1 of this topic are presented on this slide using the balanced Latin square layout shown on the previous slide. Note that the totals for the four levels of Factor A (i.e., information enhancement procedures), subjects, and presentation order are calculated in order to conduct the subsequent ANOVA on these main effects. No interactions can be assessed in this design, because these three factors are not crossed.

## 12.3.3. Testing Order Effects (Cont'd)

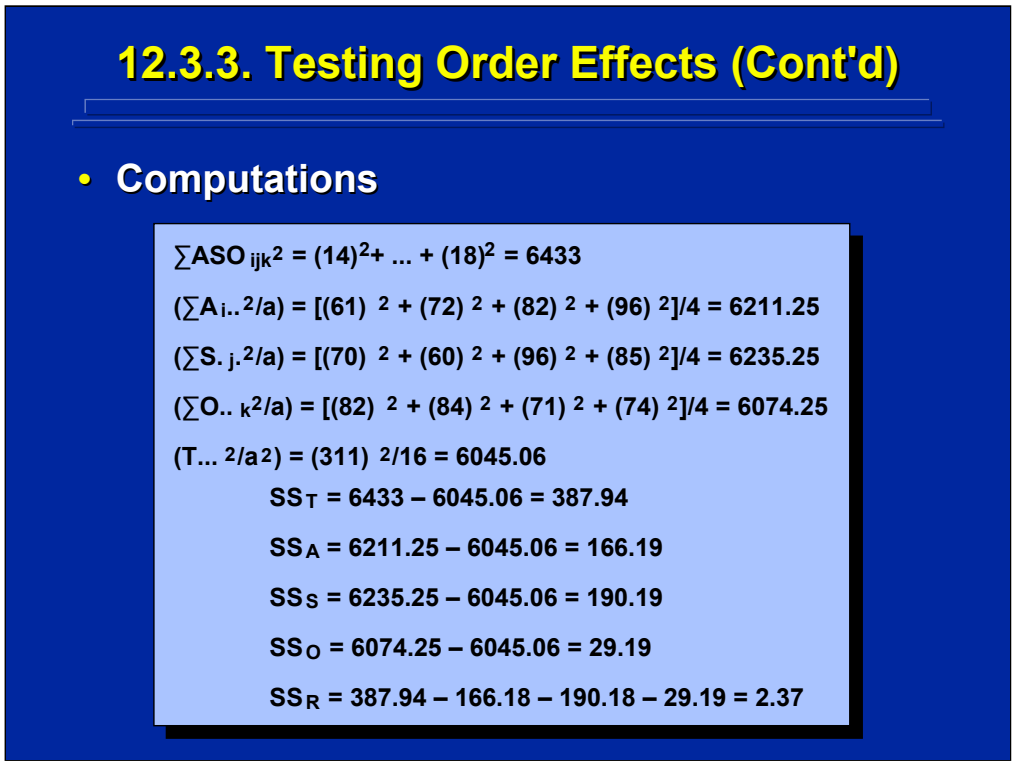
- **Computational Formulae**

| Source | df | SS |
|---|---|---|
| Total (T) | $[a^2-1]$ | $\sum ASO_{ijk}^2 - (T_{...}^2/a^2)$ |
| Factor A (A) | (a-1) | $(\sum A_{i..}^2/a) - (T_{...}^2/a^2)$ |
| Subjects (S) | (a-1) | $(\sum S_{.j.}^2/a) - (T_{...}^2/a^2)$ |
| Order (O) | (a-1) | $(\sum O_{..k}^2/a) - (T_{...}^2/a^2)$ |
| Residual (R) | (a-1)(a-2) | $SS_T - SS_A - SS_S - SS_O$ |
| | | --OR-- |
| | | $\sum ASO_{ijk}^2 - (\sum A_{i..}^2/a) - (\sum S_{.j.}^2/a) - (\sum O_{..k}^2/a) + 2(T_{...}^2/a^2)$ |

(Click in this red rectangle to see SAS calculations for this example.)

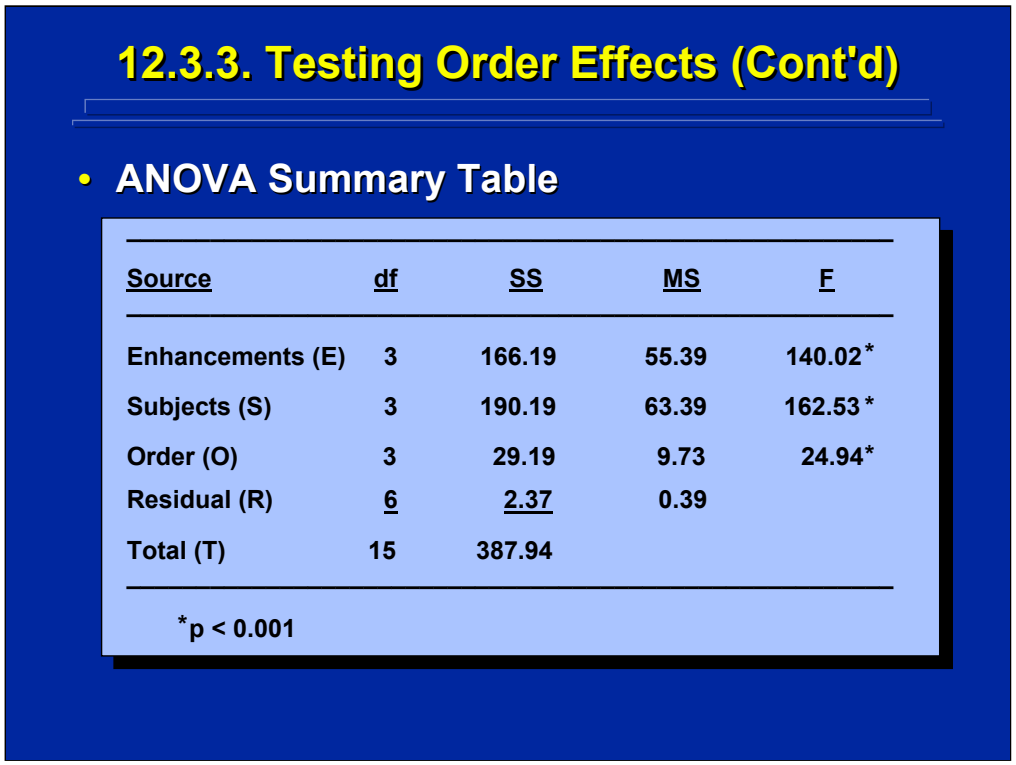
The usual algorithm for constructing SS formulae can be used to determine the total SS and the SS for the main effects of Factor A, Subjects, and Order. Since all three main effects have the same number of levels, the number of treatment levels of Factor A (i.e., a) is used to designate the df and the denominators in the SS formulae for every effect.

The error term used in the ANOVA is a pooled error term of all other variance besides the three main effects tested in the ANOVA. This error term is appropriately called Residual. The SS for Residual is simply calculated by subtraction or by using the formula presented on this slide.




(Click in this red rectangle to see SAS calculations for this example.)

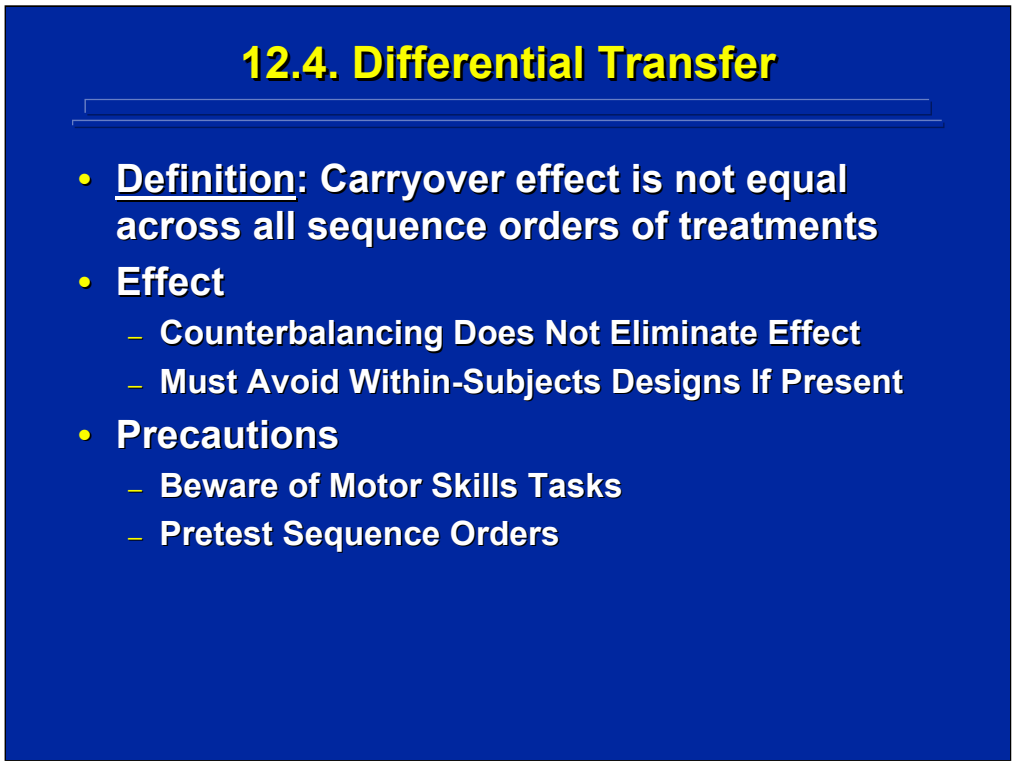
Calculations of the five component values that make up the SS values are shown on the top of this slide. The resulting SS computations for the ANOVA on the balanced Latin square design example are presented on the bottom portion of this slide.




(Click in this red rectangle to see SAS calculations for this example.)

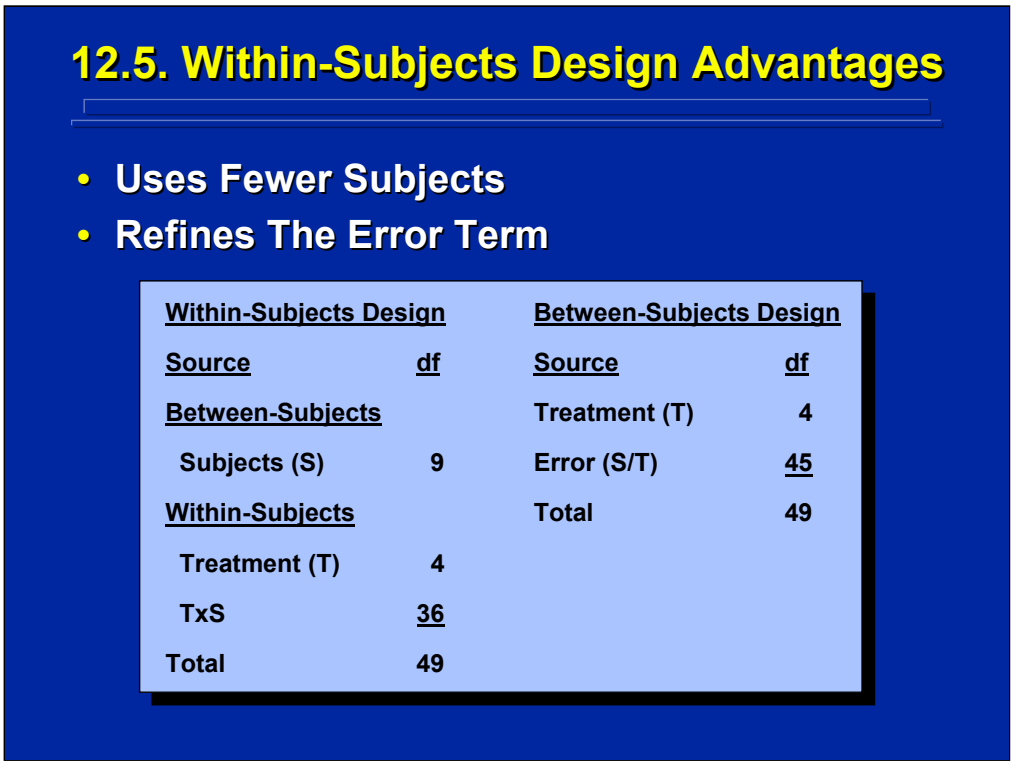
An ANOVA Summary Table on the balanced Latin square design used to balance presentation order effects in the example of a one-way, within-subjects design is presented on this slide. The main effect of Order is tested in this ANOVA. Note that the SS for Enhancements, Subjects, and Total are the same as the ANOVA of the original one-way, within-subjects design shown in Section 11.1.1. of this topic. Since Residual is used as a pooled error term in this analysis, the resulting MS and F ratios are different from the original ANOVA that used the AxS interaction as the error term.

The main effects of Enhancements, Subjects, and Order are all significant at the 0.001 level when compared to the tabled F ($F_{(3,6)} = 23.70$). Even though Order, O, is significant, is independent of the Enhancement, E, effect due to the use of the balanced Latin square and does not affect the F-test on E. Consequently, researchers usually do not test for presentation order effects and simply rely on the partial counterbalancing to protect the treatment effect from being confounded by presentation order. Note also that 24 subjects would be needed to completely counterbalance this design as compared to only four subjects used in the balanced Latin square alternative. Hence, the balanced Latin square procedure is an efficient compromise approach for partially balancing the effect of order while maintaining a small sample size.



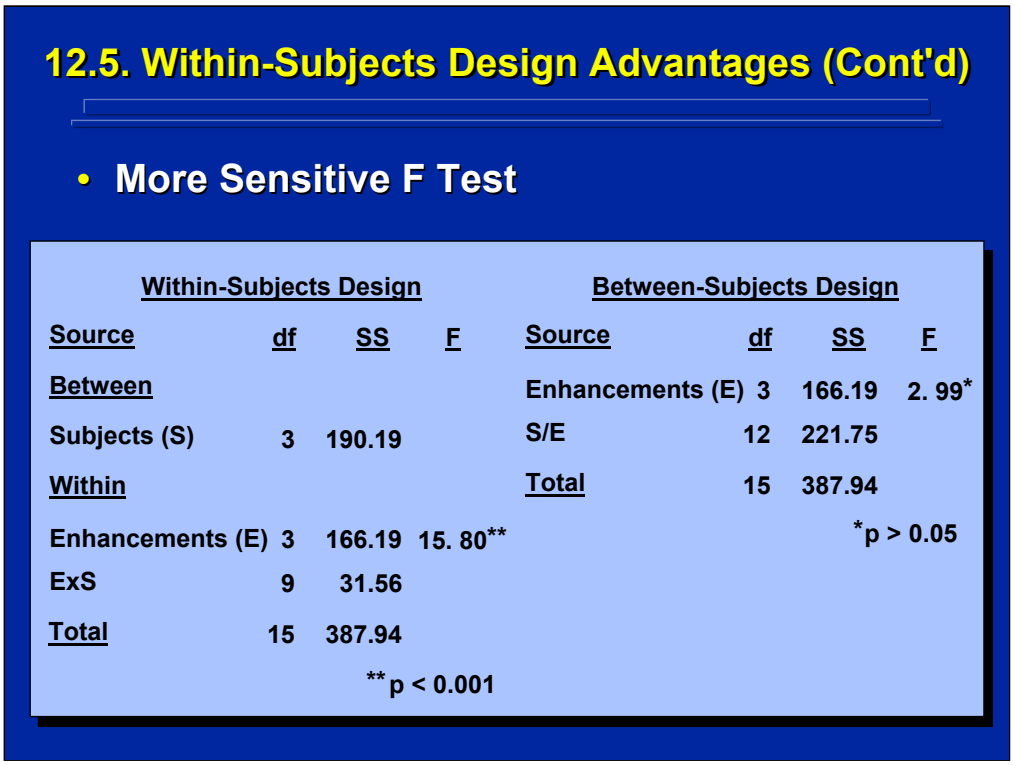

Counterbalancing is used to guard against confounding repeated measures carryover effects with the true treatment effect in within-subjects designs. Counterbalancing assumes equal carryover effects across alternative orders. When carryover effects are not equal, this situation is referred to as differential transfer (Poulton 1969). In extreme differential transfer situations, carryover only occurs when one particular level precedes another and no carryover occurs across other levels. If such differential transfer exists, it cannot be eliminated through counterbalancing. So, the experimenter should avoid using a within-subjects design and use a between-subjects design instead when marked differential transfer exists.

The experimenter should take two precautions if differential transfer is suspected. First, one should be cautious using a within-subjects design with motor skills tasks, because differential transfer frequently exists. For example, Roscoe and Williges (1975) suggested differential transfer may have affected the results of a within-subjects evaluation of aircraft attitude indicators in a flight experiment. Second, sequence order can be tested if extreme differential transfer is suspected. The treatment effects when they appear in the first position can be compared to other positions as a check on differential transfer. In any event a between-subjects design instead of a within-subjects design is the only alternative to avoid extreme differential transfer confounding.



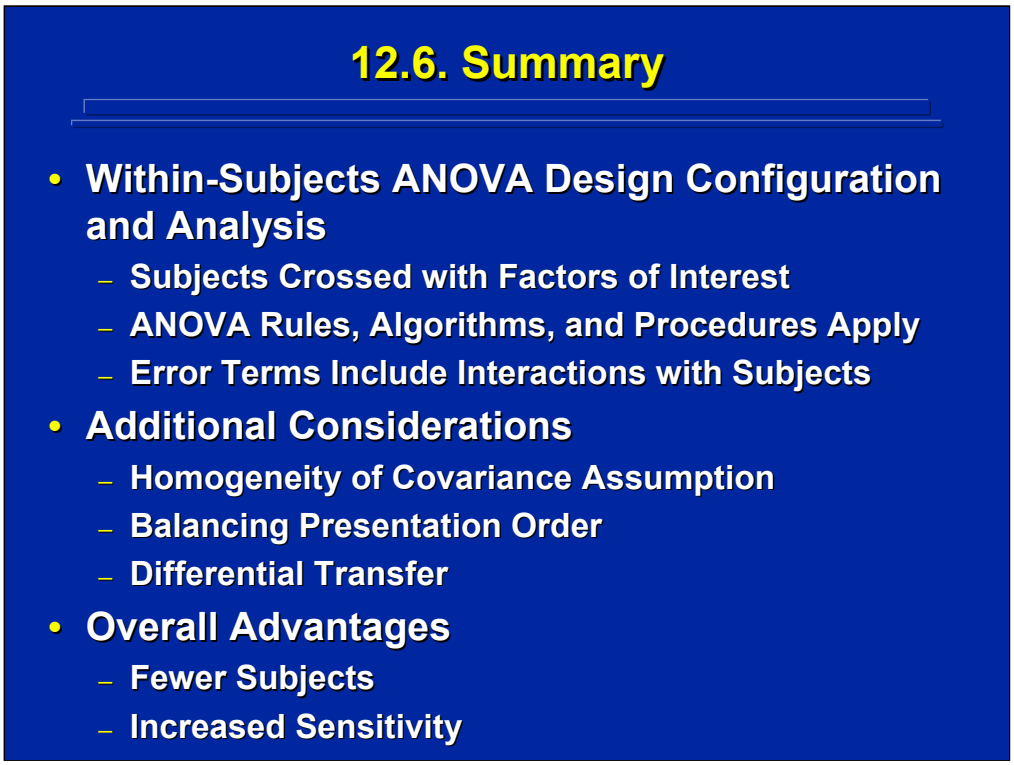

Within-subject ANOVA designs must be used when the factor of interest exists only as repeated measures. For example, practice trials and time on task are considered within-subjects variables. The experimenter, however, often considers using a within-subjects design to investigate other factors as a way of reducing the number of different subjects needed in the experiment. For example in the between-subjects and within-subjects design alternatives shown on this slide, both have a total of 50 observations. The between-subjects design requires 50 different subjects; whereas, the within-subjects design alternative requires only ten different subjects that receive all five levels of treatments where treatment order is balanced by two Balanced Latin Squares.

Within-subject designs are generally more powerful in testing an effect than a between-subjects design, because the main effect of between subject differences is removed from the error term. Difference among subjects is often the largest source of variation in a human factors experiment . The design comparison listed on this slides shows that T is tested by the TxS interaction (36 df) in the within-subjects design and the main effect of S (9 df) is removed from the error term. Alternatively, T is tested by S/T (45 df) in the between-subjects design. The large variability among subjects usually offsets the reduced df in the error term to make the within-subjects design more sensitive than its between-subjects counterpart.




(Click in this red rectangle to see SAS calculations for this example.)

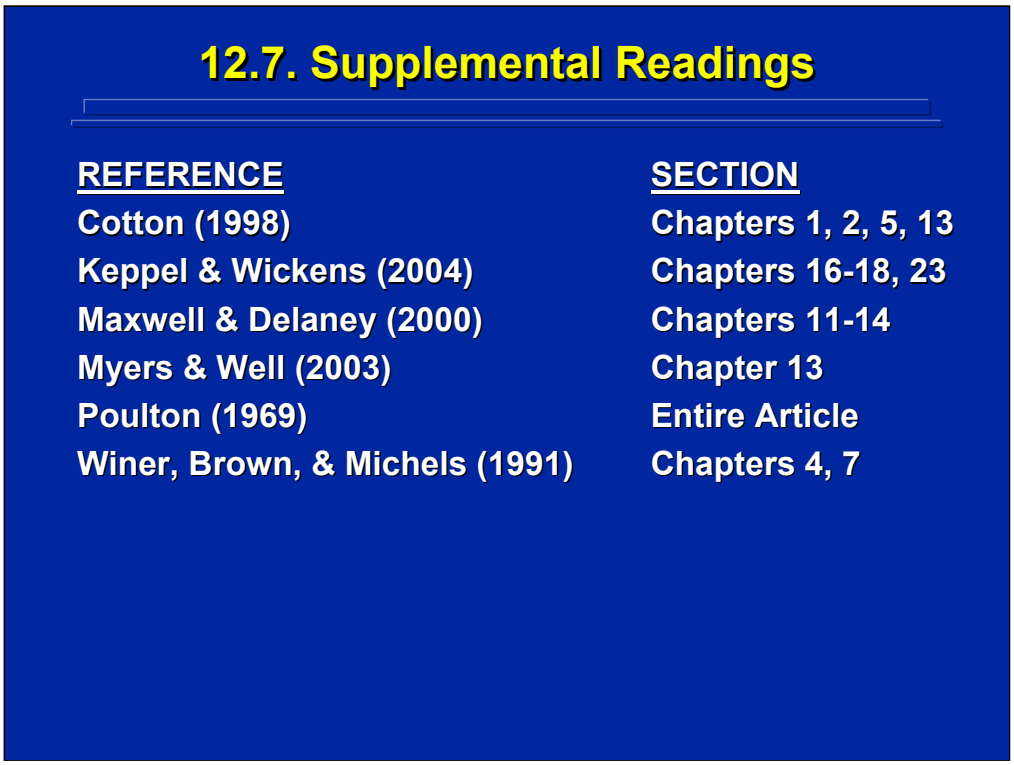
This slide provides the ANOVA Summary Table of the example problem of the one-factor, within-subjects design and its between-subjects design counterpart using the same hypothetical data set. Note that the within-subjects design used only four different subject; whereas, its between-subjects alternative would require a total of 16 different subjects (i.e., four different subjects in each of the four levels of Enhancements). The within-subjects design alternative results in a significant difference among Enhancements ($p < 0.001$), but the between-subjects alternative fails to find a significant difference ($p > 0.05$). Even though the between-subjects design alternative has more degrees of freedom in the error term than the within-subjects design (i.e. 15 df versus 9 df), the pooled $SS_{Error}$ of the between-subjects alternative (221.75) is much larger than the $SS_{Error}$ of the within-subjects alternative (31.56) that removes the SS of the main effect of subjects (190.19) from the error term. Hence, the within-subjects design alternative requires fewer subjects and provides a more sensitive F-test than its between-subjects design counterpart.




By way of summary, this topic covered within-subjects ANOVA design configurations and analyses that require subjects to be crossed with all the factors of interest in the experiment. If the researcher chooses a within-subjects design, all the ANOVA rules, algorithms and procedures apply. Assuming all factors of interest are fixed-effects factor, the interaction of factor(s) with subjects is the appropriate error term for tests of significance in these repeated measures designs.

Additionally, the experimenter must consider the assumption of homogeneity of covariance and make adjustments to the F table if marked deviations from circularity are expected. Since every subject receives every treatment in a within-subjects design, the researcher also needs to balance presentation order through complete counterbalancing, partially balanced Latin squares, or random assignment procedures. Balanced Latin square procedures are most useful in human factors research and have implications for choice of sample size. Marked carryover effects as demonstrated be differential transfer may preclude the use of within-subjects designs.

Overall, the within-subjects design is more sensitive and requires fewer subjects than its between-subjects counterpart. Some variables such as practice only exist as repeated measures. Others such as type of training cannot be manipulated as a within-subjects factor. Many factors, however, can be investigated as either within-subjects or between-subjects factors.




Within-subject designs are commonly used in behavioral science research. Appropriate chapters in common experimental design textbooks used by human factors researchers are listed on this slide. All of these texts cover univariate approaches to within-subjects designs similar to the presentation in this reference material. The most extensive discussion to multivariate approaches to within-subjects designs is covered in the supplemental reading by Maxwell and Dulaney (2000). In Chapter 5, Cotton (1998) describes SAS general linear model (GLM) analytical procedures for testing the overall order effect as well as other various carryover effects across orders in factorial ANOVA designs. Additionally, the Cotton (1998) reference discusses special purpose crossover design alternatives, and the Poulton (1969) article discusses differential transfer.

# Topic 13. Mixed-Factors ANOVA Designs

13.1. Mixed-Factors Design Configurations
- 13.1.1. Two-Factor Design
- 13.1.2. Two-Factor Design Example
- 13.1.3. Three-Factor Design
- 13.1.4. n-Factor Design

13.2. Mixed-Factors Design Considerations

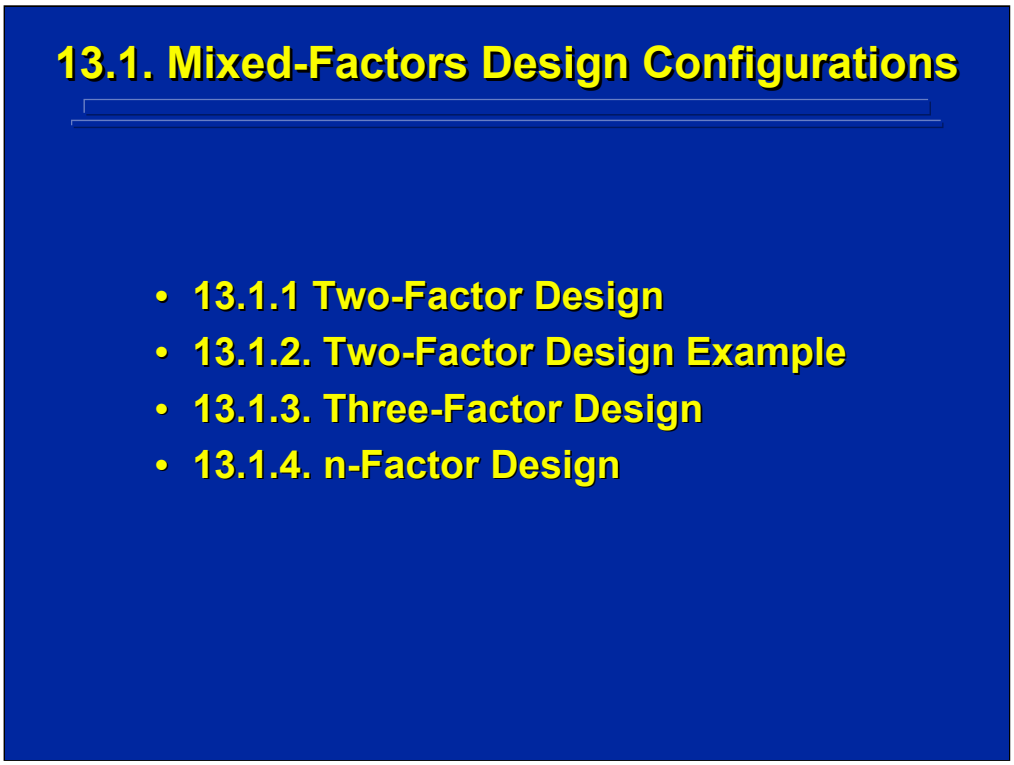
13.3. Summary

13.4. Supplemental Readings

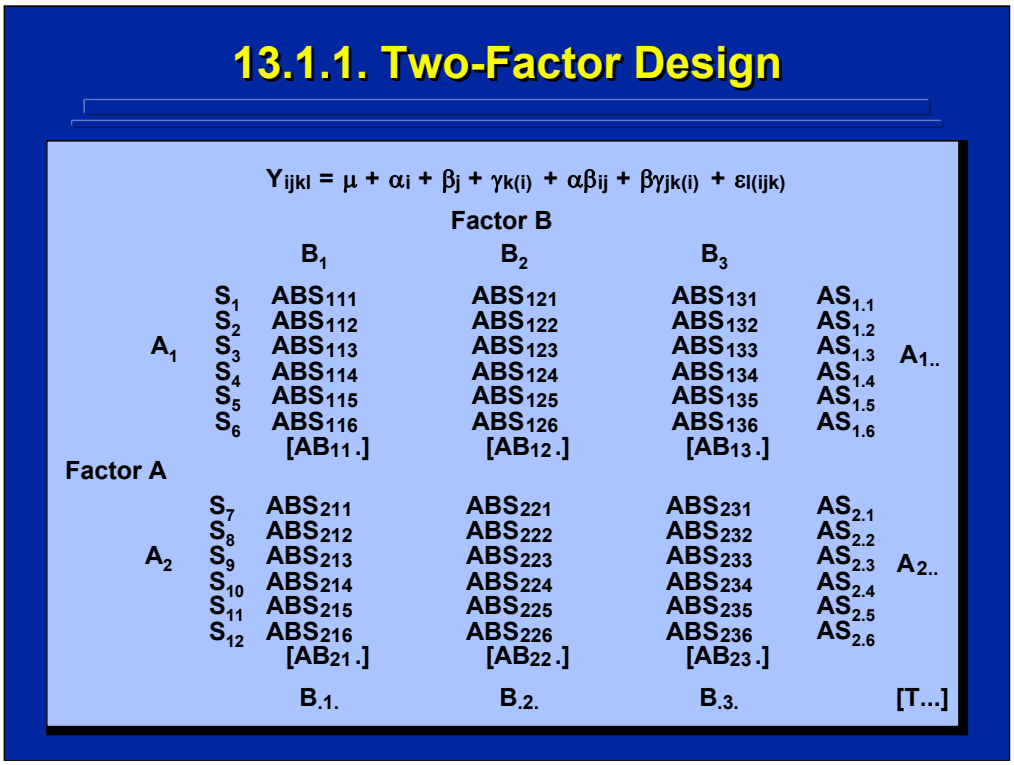
This topic covers the basic configuration and analytical procedures used in mixed-factors ANOVA designs which comprise the third major category of ANOVA designs. Mixed-factors designs are composed of both between-subjects and within-subjects factors. This type of ANOVA design is often referred to as split-plot designs in the scientific literature.

These designs are used quite frequently in human factors and ergonomic research due to the nature of the independent variables being investigated in the experiment. Consider a training research study that investigates training methods and practice trials. The researcher must manipulate the training condition variable as a between-subjects variable because subjects cannot return to a beginning level of knowledge when provided with alternative training. On the other hand, practice trials in this training experiment must be manipulated as a within-subjects factor since each subject receives multiple trials. Consequently, a mixed-factors design is needed.

This topic also provides a list of considerations that the research should address when using mixed-factors designs and ends with a list of recommended supplemental readings in experimental design textbooks dealing with mixed-factors designs.

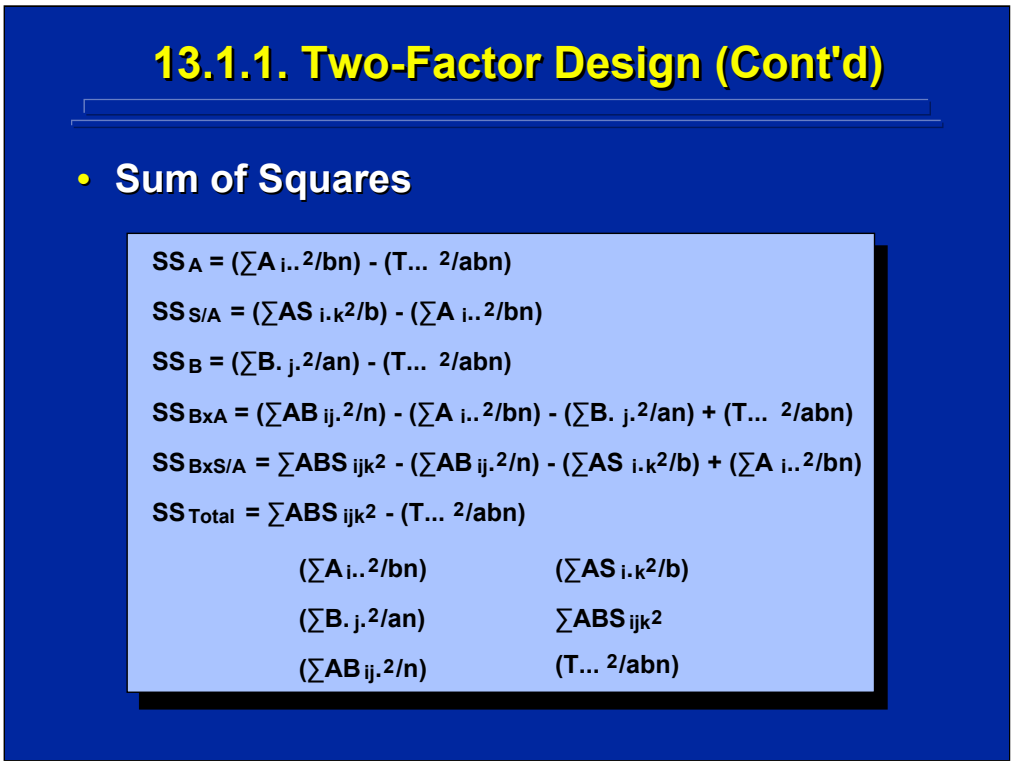



A mixed-factors ANOVA design, by definition, must have a minimum of one between-subjects factor and one within-subjects factor. Consequently, a two-factor design is the smallest possible mixed-factors design. After a two-factor design is described in terms of the simplified notation, a computational example is provided. Both three-factor and generalizations to n-factor designs are described to emphasize that all the general procedures, rules, and algorithms for ANOVA designs using the simplified notation also apply to any factorial mixed-factors design.

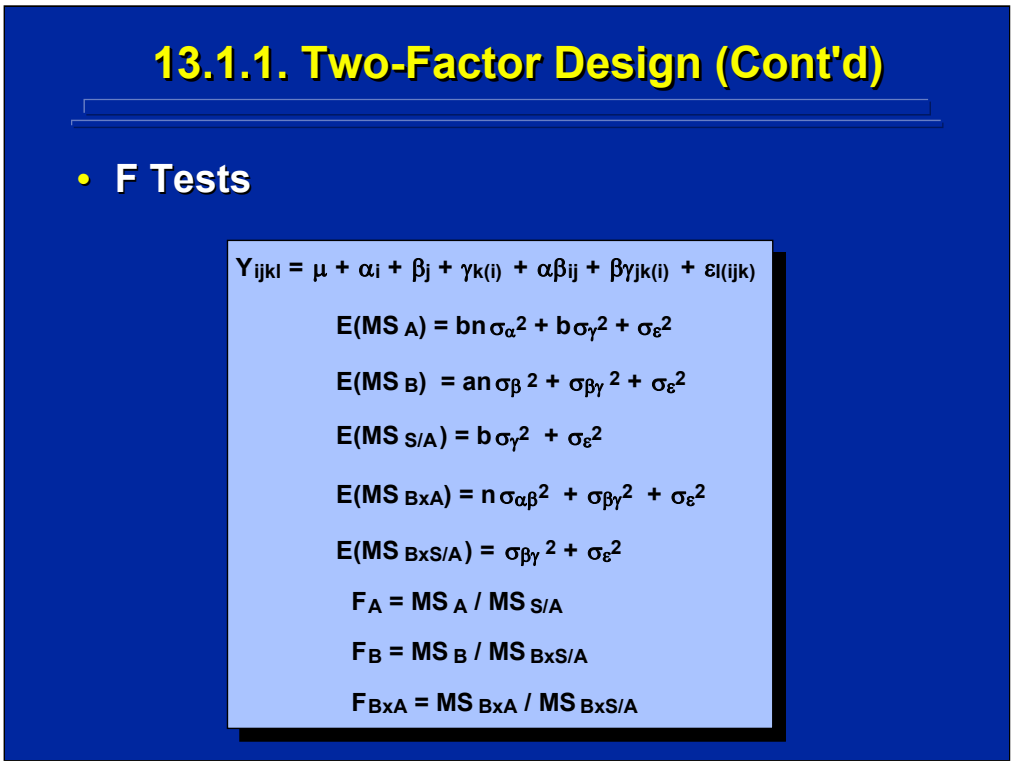



This slide lists the statistical model and the data matrix layout in the simplified notation for the basic two-way, mixed-factors ANOVA design. Note that the statistical model lists subjects being nested in Factor A and crossed with Factor B. Consequently, the data matrix shows that when n = 6, the six levels of subjects in $A_1$ are different than the six levels of subjects in $A_2$ resulting in a total of 12 different subjects required for participation in the experiment. Each of these 12 subjects receives all three levels of the within-subjects factor, Factor B. The choice of six subjects per cell is appropriate in order to completely counterbalance the three levels of the within-subjects Factor B.

If the statistical model were changed such that subjects were crossed with Factor A and nested in Factor B, then a total of 18 different subjects would be required if cell size remained 6. And, subject designation in the data matrix would be changed accordingly.




The experimenter can use the SS algorithm for generating the SS formulae for mixed-factors designs. The top portion of this slide shows the various SS computational formulae for a two-way, mixed-factors design using the simplified notation. Note that the formulae are made up of various combinations of the six component values listed at the bottom of this slide. Each of these component values are listed on the previous slide.




The experimenter can use the algorithm for specifying the expected mean squares based on the statistical model of the two-way, mixed-factors design. The resulting E(MS) for this design are listed on this slide.

By using the rules for generating F ratios, one can determine that $MS_{S/A}$ is the appropriate error term for testing Factor A, and $MS_{BxS/A}$ is the appropriate error term for testing both Factor B and the AxB interaction. These three resulting F ratios are shown on the bottom portion of this slide.

## 13.1.1. Two-Factor Design (Cont'd)

- **ANOVA Summary Table**

| Source | df | SS | MS | F |
|---|---|---|---|---|
| **Between** | | | | |
| A | a-1 | $SS_A$ | $MS_A$ | $MS_A/MS_{S/A}$ |
| S/A | a(n-1) | $SS_{S/A}$ | $MS_{S/A}$ | |
| **Within** | | | | |
| B | b-1 | $SS_B$ | $MS_B$ | $MS_B/MS_{BxS/A}$ |
| BxA | (a-1)(b-1) | $SS_{BxA}$ | $MS_{BxA}$ | $MS_{BxA}/MS_{BxS/A}$ |
| BxS/A | a(b-1)(n-1) | $SS_{BxS/A}$ | $MS_{BxS/A}$ | |
| **Total** | abn-1 | $SS_{total}$ | | |

The general format for specifying the ANOVA Summary Table is shown on this slide for the two-factor design. Both between-subjects and within-subjects effects are listed for mixed-factors designs. Note that Factor A is the between-subjects factor and Factor B is the within-subjects factor as previously specified in the statistical model. The error terms are grouped with the effects being tested. Based on the E(MS) shown on the previous page, the S/A error term is grouped with A as a between-subjects effect, and the BxS/A error term is grouped with both B and BxA as a within-subjects effects.

## 13.1.2. Two-Factor Design Example

- **Example Problem: The decrement in target detection across 1-hour monitoring sessions was measured every 20 minutes for five soldiers who monitored displays where the ratio of targets to non-targets was either 9/1 or 1/9. Are there any significant effects ($p < 0.05$) in the percent of defined targets detected in this experiment?**

(Click in this red rectangle to see SAS calculations for this example.)

This slide describes a two-way, mixed-factors example problem that has a sample size of 5 (i.e., n = 5). The ratio of targets to non-targets is treated as a between-subjects factor and has two levels, 9/1 or 1/9. The three successive 20-minute monitoring sessions are, by definition, levels of a within-subjects factor because each subject must participate in each of the three successive sessions during the 1-hour monitoring period. The Slater and Williges (2006) appendix describe the SAS analysis for this example problem of a mixed-factors design.

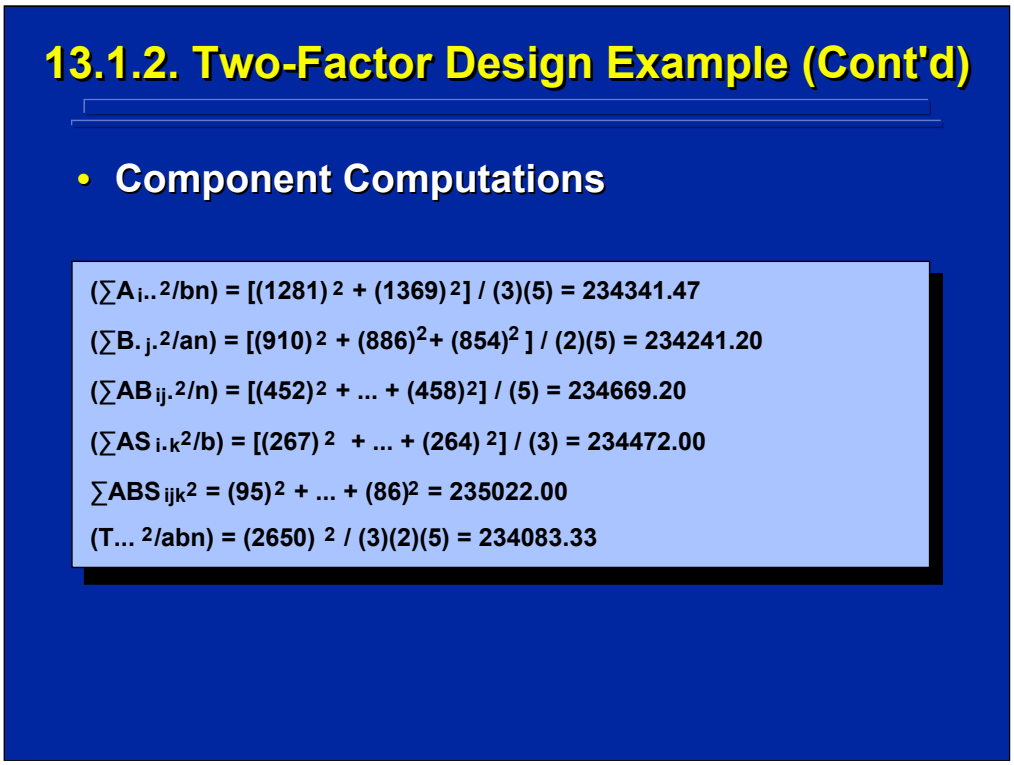
## 13.1.2. Two-Factor Design Example (Cont'd)

| S/N Ratio (Factor A) | | Time Monitoring (Factor B) 0-20 min. | 20-40 min. | 40-60 min. | | |
|---|---|---|---|---|---|---|
| 1/9 | $S_1$ | 95 | 90 | 82 | $AS_{1.1} = 267$ | |
| | $S_2$ | 89 | 82 | 83 | $AS_{1.2} = 254$ | |
| | $S_3$ | 92 | 80 | 79 | $AS_{1.3} = 251$ | $A_{1..} = 1281$ |
| | $S_4$ | 86 | 89 | 77 | $AS_{1.4} = 252$ | |
| | $S_5$ | 90 | 92 | 75 | $AS_{1.5} = 257$ | |
| | | [452] | [433] | [396] | | |
| 9/1 | $S_6$ | 90 | 88 | 92 | $AS_{2.1} = 270$ | |
| | $S_7$ | 87 | 95 | 95 | $AS_{2.2} = 277$ | |
| | $S_8$ | 96 | 93 | 95 | $AS_{2.3} = 284$ | $A_{2..} = 1369$ |
| | $S_9$ | 94 | 90 | 90 | $AS_{2.4} = 274$ | |
| | $S_{10}$ | 91 | 87 | 86 | $AS_{2.5} = 264$ | |
| | | [458] | [453] | [458] | | |
| | | $B_{.1.} = 910$ | $B_{.2.} = 886$ | $B_{.3.} = 854$ | | [T... = 2650] |

(Click in this red rectangle to see SAS calculations for this example.)

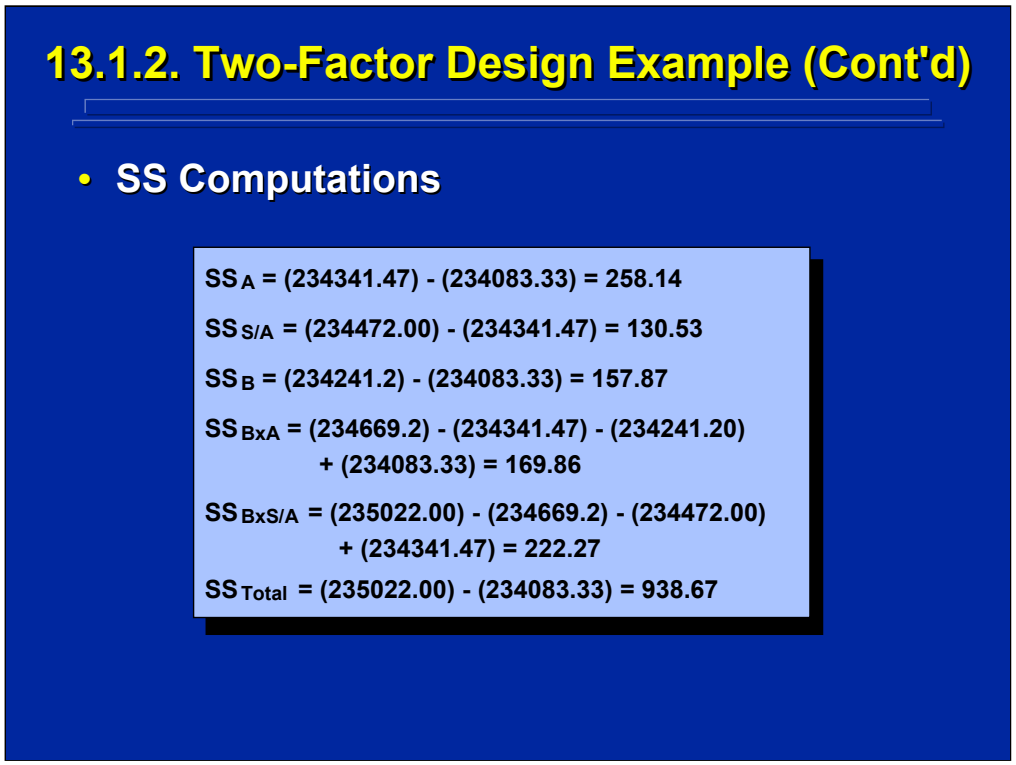
Hypothetical target detection percentage data for the example problem are presented in the data matrix shown on this slide. The simplified notation is used to show various totals used in the subsequent SS calculations. Note that the 30 detection probabilities shown across the six cells of the design are designated as various $ABS_{ijk}$ scores and the cell totals shown in brackets are various $AB_{ij.}$ totals.

Note that the three levels of the within-subjects factor, Time Monitoring, cannot be counterbalanced because the three 20-minute sessions can only occur successively. Consequently, a sample size (i.e., n = 5) was chosen by the experimenter without concern for counterbalancing. If counterbalancing were possible, then a sample size of six would be more appropriate in order to completely counterbalance the three levels of the within-subjects factor in the mixed-factors design.




(Click in this red rectangle to see SAS calculations for this example.)

The six component computations of the example data that are used in the SS calculations for this mixed-factors design example are shown on this slide.




(Click in this red rectangle to see SAS calculations for this example.)

The experimenter can use the components scores calculated on the previous slide to determine the various SS values from the SS formulae stated in the simplified notation for the mixed-factors design example.

## 13.1.2. Two-Factor Design Example (Cont'd)

- **ANOVA Summary Table**

| Source | df | SS | MS | F |
|---|---|---|---|---|
| **Between** | | | | |
| Ratio (R) | 1 | 258.14 | 258.14 | 15.82** |
| S/R | 8 | 130.53 | 16.32 | |
| **Within** | | | | |
| Time (T) | 2 | 157.87 | 78.94 | 5.68* |
| TxR | 2 | 169.86 | 84.93 | 6.11* |
| TxS/R | 16 | 222.27 | 13.89 | |
| **Total** | 29 | 938.67 | | |

*p < 0.05 **p < 0.01

(Click in this red rectangle to see SAS calculations for this example.)

The complete Summary Table for the mixed-factors design example is shown on this slide. The Summary Table is presented in the standard format using designations for the actual factors manipulated in the experiment rather than generic Factor A and B designations.

Note that two of the F-ratios indicate significant differences at the 0.05 level and one indicates significance at the 0.01 when compared to F tabled values. Since Ratio only has two levels, the experimenter can conclude that soldiers detected significantly more targets overall in the 9/1 ratio condition than in the 1/9 ratio condition. Additional post hoc comparisons are needed to isolate the main effect of Time and the RxT interaction. Looking at the cell means in the data set, it appears that decrease in detection probability across the successive 20-minute monitoring periods appear to be restricted to the 1/9 ratio condition that is characteristic of classical vigilance decrements in human factors research. In fact, Williges (1969) reported similar results in an actual monitoring experiment and interpreted these results in terms of signal detection theory.

## 13.1.3. Three-Factor Design

$$Y_{ijklm} = \mu + \alpha_i + \beta_j + \delta_k + \gamma_{l(ij)} + \alpha\beta_{ij} + \alpha\delta_{ik} + \beta\delta_{jk} + \delta\gamma_{kl(ij)} + \alpha\beta\delta_{ijk} + \varepsilon_{m(ijkl)}$$

| Source | df | SS | MS | F |
|---|---|---|---|---|
| **Between** | | | | |
| A | a-1 | $SS_A$ | $MS_A$ | $MS_A/MS_{S/AB}$ |
| B | b-1 | $SS_B$ | $MS_B$ | $MS_B/MS_{S/AB}$ |
| AxB | (a-1)(b-1) | $SS_{AxB}$ | $MS_{AxB}$ | $MS_{AxB}/MS_{S/AB}$ |
| S/AB | ab(n-1) | $SS_{S/AB}$ | $MS_{S/AB}$ | |
| **Within** | | | | |
| C | c-1 | $SS_C$ | $MS_C$ | $MS_C/MS_{CxS/AB}$ |
| AxC | (a-1)(c-1) | $SS_{AxC}$ | $MS_{AxC}$ | $MS_{AxC}/MS_{CxS/AB}$ |
| BxC | (b-1)(c-1) | $SS_{BxC}$ | $MS_{BxC}$ | $MS_{BxC}/MS_{CxS/AB}$ |
| AxBxC | (a-1)(b-1)(c-1) | $SS_{AxBxC}$ | $MS_{AxBxC}$ | $MS_{AxBxC}/MS_{CxS/AB}$ |
| CxS/AB | ab(c-1)(n-1) | $SS_{CxS/AB}$ | $MS_{CxS/AB}$ | |
| **Total** | abcn-1 | $SS_{total}$ | | |

All the general procedures, rules, and algorithms apply to higher-order, mixed-factors designs. Always begin by stating the statistical model of the design. This slide shows the ANOVA Summary Table of a three-way, mixed-factors design that includes two between-subjects factors, A and B, and one within-subjects factor, C. The statistical model shows this designation in the nesting relationship for subjects, $\gamma$.

Notice that the between-subjects effects are grouped with their single error term, S/AB. Likewise, the within-subjects effects are grouped with their single error term, CxS/AB.

## 13.1.3. Three-Factor Design (Cont'd)

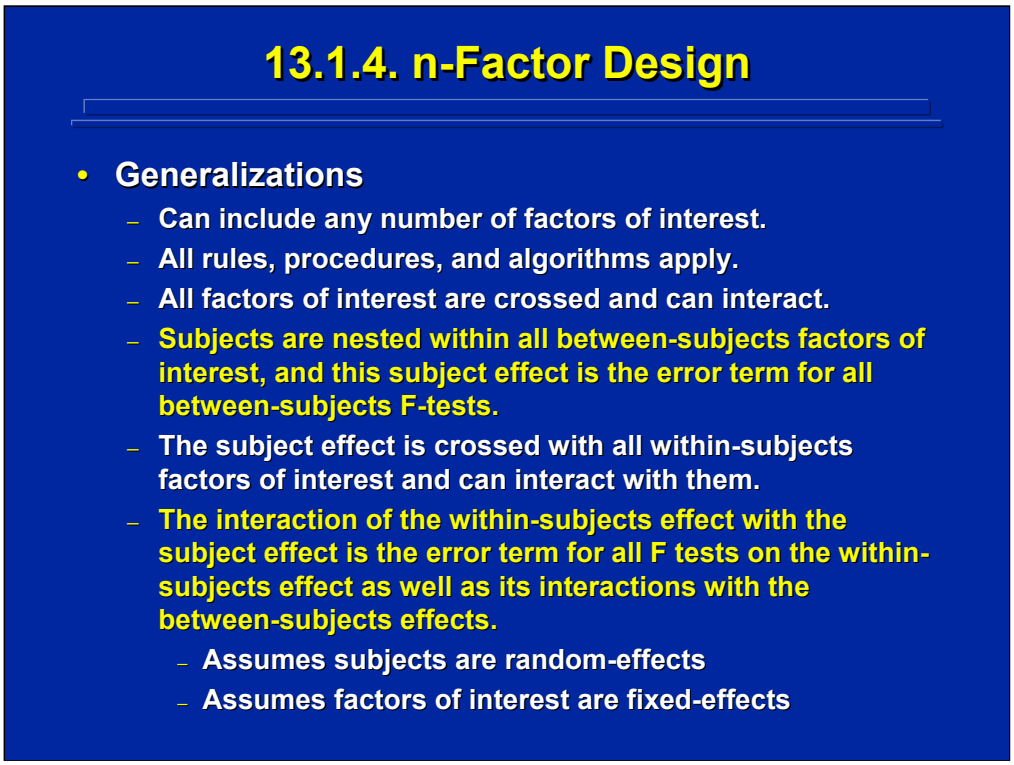
$$Y_{ijklm} = \mu + \alpha_i + \beta_j + \delta_k + \gamma_{l(i)} + \alpha\beta_{ij} + \alpha\delta_{ik} + \beta\delta_{jk} + \beta\gamma_{jl(i)} + \delta\gamma_{kl(i)} + \alpha\beta\delta_{ijk} + \beta\delta\gamma_{jkl(i)} + \varepsilon_{m(ijkl)}$$

| Source | df | SS | MS | F |
|---|---|---|---|---|
| **Between** | | | | |
| A | a-1 | $SS_A$ | $MS_A$ | $MS_A/MS_{S/A}$ |
| S/A | a(n-1) | $SS_{S/A}$ | $MS_{S/A}$ | |
| **Within** | | | | |
| B | b-1 | $SS_B$ | $MS_B$ | $MS_B/MS_{BxS/A}$ |
| AxB | (a-1)(b-1) | $SS_{AxB}$ | $MS_{AxB}$ | $MS_{AxB}/MS_{BxS/A}$ |
| BxS/A | a(b-1)(n-1) | $SS_{BxS/A}$ | $MS_{BxS/A}$ | |
| C | c-1 | $SS_C$ | $MS_C$ | $MS_C/MS_{CxS/A}$ |
| AxC | (a-1)(c-1) | $SS_{AxC}$ | $MS_{AxC}$ | $MS_{AxC}/MS_{CxS/A}$ |
| CxS/A | a(c-1)(n-1) | $SS_{CxS/A}$ | $MS_{CxS/A}$ | |
| BxC | (b-1)(c-1) | $SS_{BxC}$ | $MS_{BxC}$ | $MS_{BxC}/MS_{BxCxS/A}$ |
| AxBxC | (a-1)(b-1)(c-1) | $SS_{AxBxC}$ | $MS_{AxBxC}$ | $MS_{AxBxC}/MS_{BxCxS/A}$ |
| BxCxS/A | a(b-1)(c-1)(n-1) | $SS_{BxCxS/A}$ | $MS_{BxCxS/A}$ | |
| **Total** | abcn-1 | $SS_{total}$ | | |

This slide shows the ANOVA Summary Table of an alternative three-way, mixed-factors design that includes only one between-subjects factor, A, and two within-subjects factor, B and C. The statistical model shows this designation in the nesting relationship for subjects, $\gamma$. Again, the experimenter should always begin by stating the statistical model of the design in order to determine all the effects that can be estimated.

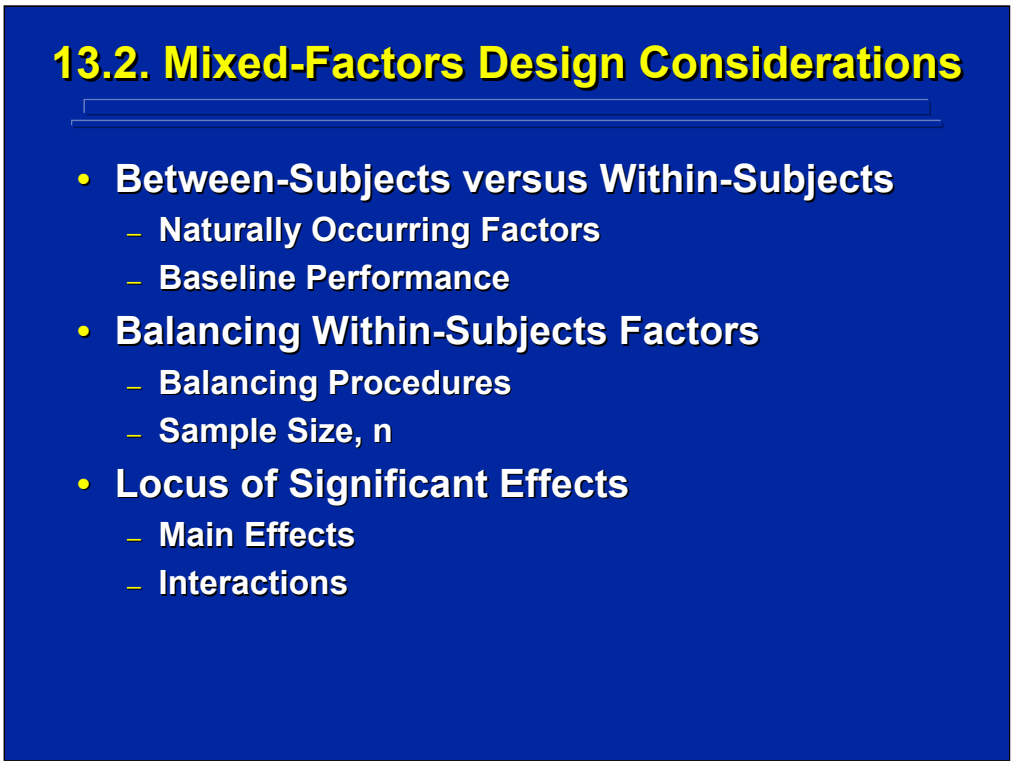
Notice that the single between-subjects effect, A, is grouped with its error term, S/A. Likewise, the within-subjects effects are grouped with their error terms. One can determine that there are three possible within-subjects error terms, BxS/A, CxS/A, and BxCxS/A, based on the E(MS) designations for this mixed-factors design. The three groupings of effects with their error term are shown on this slide.

The same main effects, two-way interactions, and the three-way interaction tested in the previous three-way design example are tested in this design, but different error terms are used due to the crossed and nesting relationship of factors. The experimenter can always determine the appropriate relationship for any mixed-factors design by following the general ANOVA procedures, rules, and algorithms.



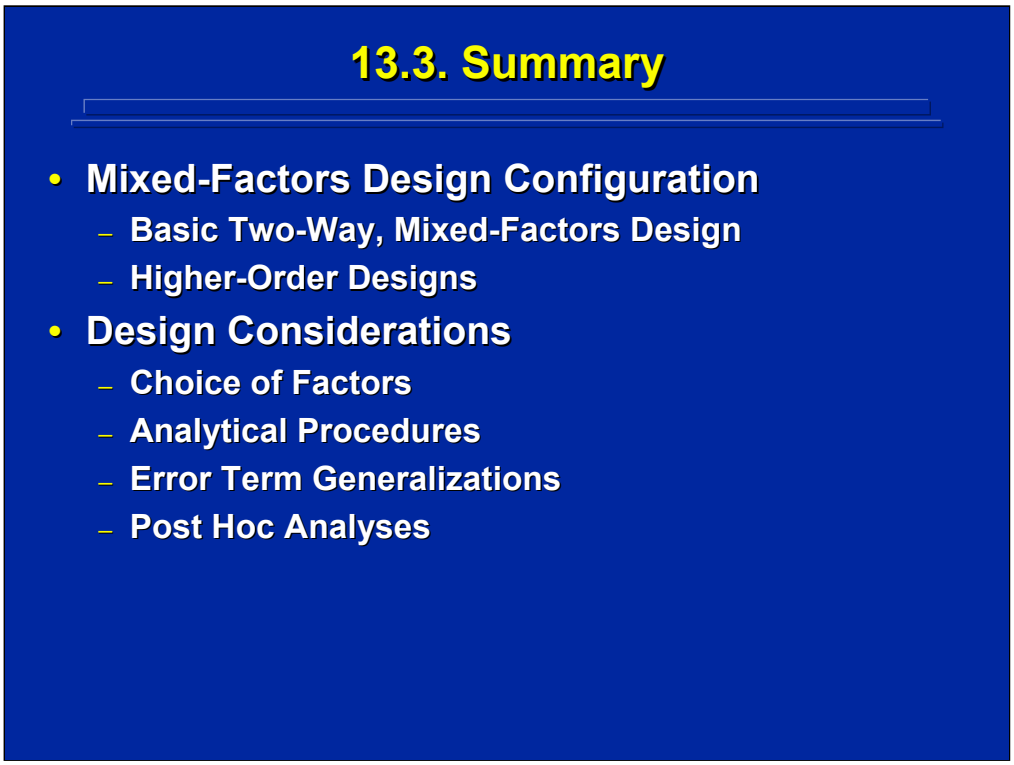

This slide provides generalizations for any mixed-factors design that has equal sample size and all the factors of interest are crossed and considered fixed-effects variables. The experimenter should begin by stating the statistical model of the mixed-factors design and then follow all the ANOVA rules and algorithms for stating the sources of variation, SS formulae, E(MS) values, and possible F-ratios.

Notice that the error terms of the between-subjects effects are simply subjects nested within those effects. Likewise, the error terms for within-subjects effects are the interactions of those effects with the nested subject effect. The regular grouping procedures for between-subjects and within-subjects effects are used in specifying the resulting ANOVA Summary Table.



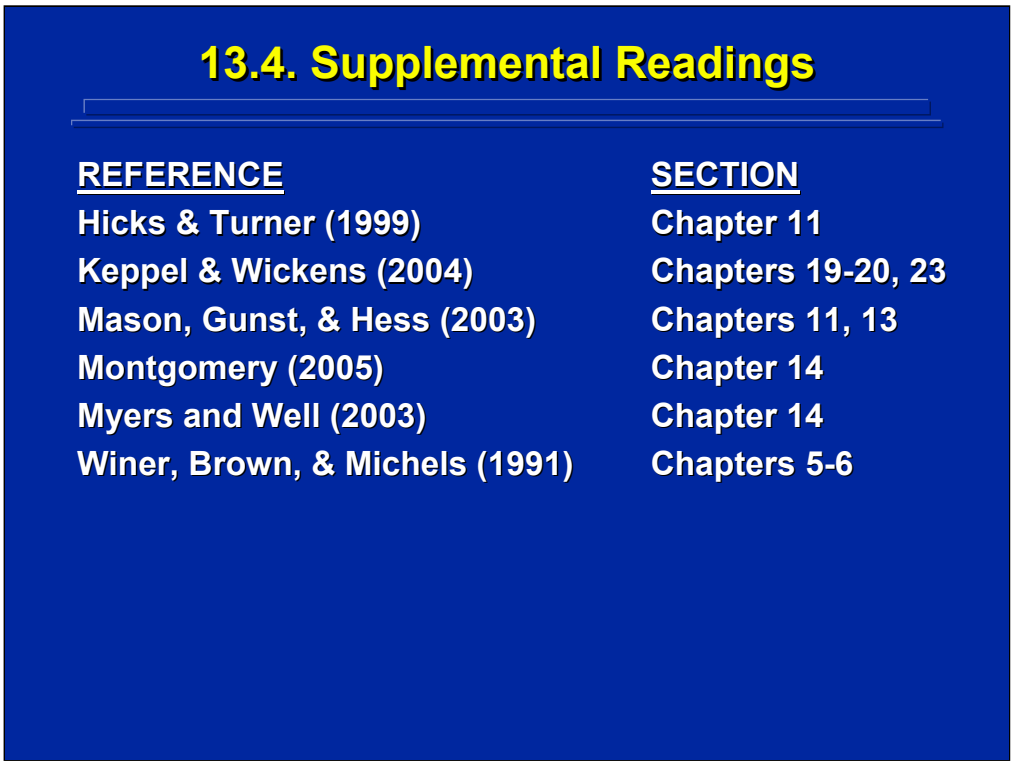

The experimenter must decide which factors are manipulated as between-subjects factors and which factors are treated as within-subjects factors when using a mixed-factors design. Usually this decision is determined by the natural occurrence of the factors. For example, Williges, Johnston, and Briggs (1967) used a three-way mixed-factors transfer of training design to investigate verbal communication in teamwork. Exposure to the various communication conditions during training and transfer required between-subjects manipulation, but the practice trials factor was naturally a within-subjects factor. If independent variables are manipulated as within-subjects factors, care must be taken that human performance is not affected by repeated measures and these factors are measured at the same baseline of performance as would occur with between-subjects effects.

The within-subjects factors in a mixed-factors design must be balanced to control order and sequence effects as discussed in Topic 12 dealing with within-subjects designs. Usually, a Balanced Latin Square is the most efficient balancing procedure requiring the minimum number of subjects. The overall results of F-tests in mixed-factors designs involving more than two-levels of a factor require additional post hoc tests as described in Topic 11 in order to isolate the locus of main effects and interactions.




Mixed-factors designs are one of the most useful ANOVA designs for human factors, because researchers often need to consider variables simultaneously that naturally appear as between-subjects and within-subjects variables in the same experiment. The basic mixed-factors design involves two factors (i.e., one between-subjects factor and one within-subjects factor). The experimenter can generate any higher-order, mixed-factors, factorial design with more than two factors involving any combination of between-subjects and within-subjects factors.

The experimenter must decide which variables are manipulated as between-subjects variables and which variables are manipulated as within-subjects variables in a mixed-factors design. The nature of the real-world variable is the overriding parameter in making this decision. Gender, for example, would always be a between-subjects factor. Carryover effects that could change baseline performance must be considered with all within-subjects factors. All data analysis procedures used in mixed-factors designs such as calculating the overall ANOVA, choosing error terms for F-tests, and conducting post hoc analyses on main effects and interactions follow the same rules and algorithms used for between-subjects and within-subjects ANOVA designs once the appropriate mixed-factors statistical model is stated.

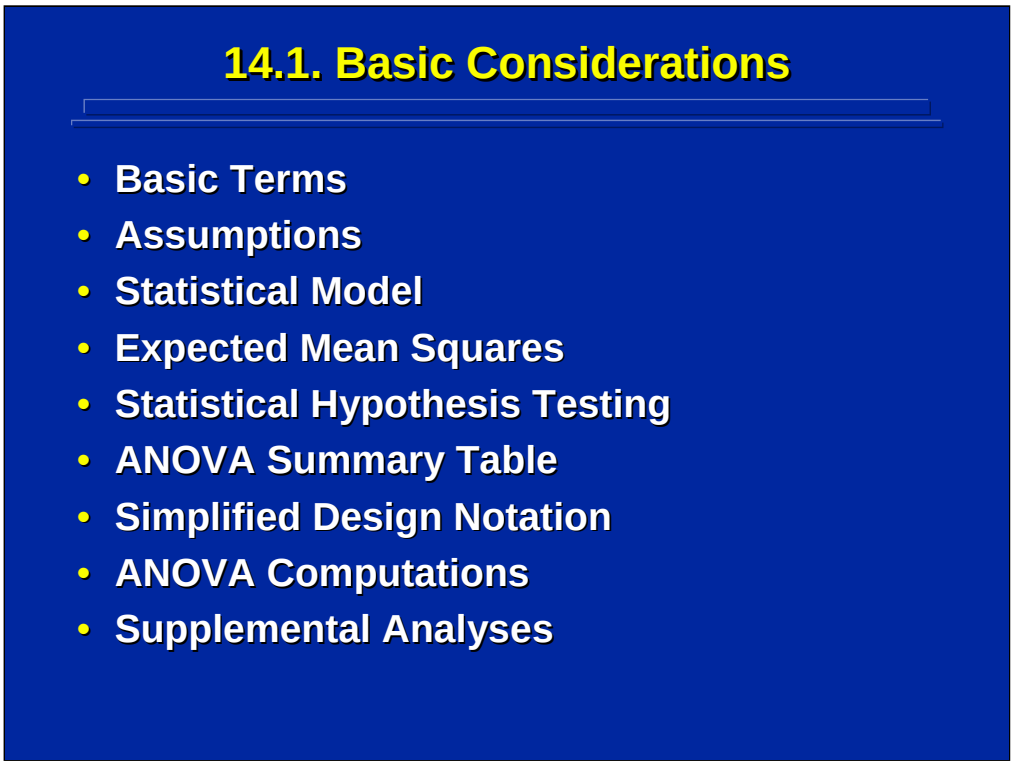



Appropriate chapters dealing with mixed-factors or split-plot designs in common experimental design textbooks used by human factors researchers are listed on this slide. The chapters in Keppel and Wickens (2004) and Winer et al. (1991) most closely follow the procedures covered in this topic.

# Topic 14. Summary of Basic ANOVA

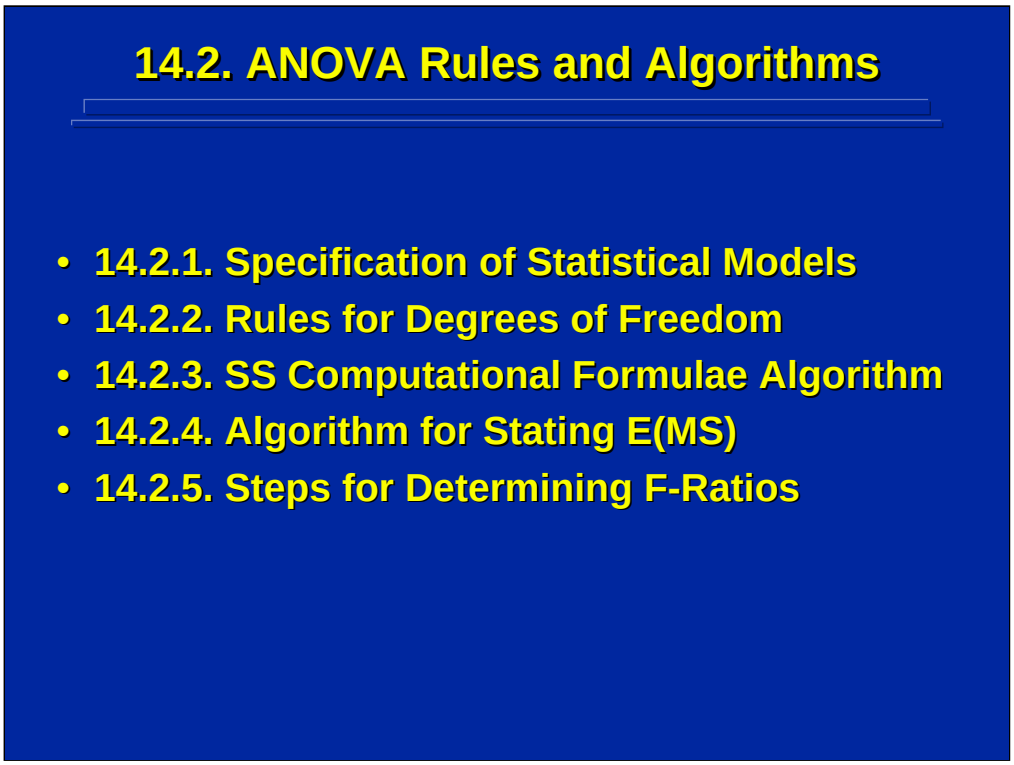
14.1. Basic Considerations
14.2. ANOVA Rules and Algorithms
14.3. Design Classification
14.4. n-Factor Design Generalizations
14.5. ANOVA Design Process
14.6. Summary
14.7. Supplemental Readings

This topic provides a brief summary review and roadmap of basic ANOVA in terms of fundamental considerations, ANOVA design classification, generalizations, steps in the overall ANOVA design and data interpretation process. Details on these issues are provided in the topics referenced. This topic ends with a summary of recommended supplemental readings in experimental design textbooks dealing with basic ANOVA.




Every ANOVA design can be described using a standard terminology including factors, factor level, crossed and nested factors, factorial design, treatment cell, interaction and factors can be identified as fixed or random. Knowing these terms and the basic assumptions of homogeneity of variance and normally distributed variables as described in Topic 8 are central to understanding ANOVA. The additional assumption of homogeneity of covariance must be considered for within-subjects designs as described in Topic 12. Usually an equal sample size is used in human factors research to provide robustness against basic assumption violations.

Using standard rules, procedures and algorithms, the researcher begins by stating the statistical model of the ANOVA design in order to determine the effects of interest that can be estimated from the experiment. Subsequently, the error terms to be used in statistical hypothesis testing using F-ratios are determined through the expected mean squares. The results of any ANOVA can be summarized in a Summary Table that provides sources of variation, degrees of freedom, sum of squares, mean squares, and F-ratios. Conventions are followed for grouping effects in ANOVA Summary Tables, and every F-test can be specified in a standard format as presented in Topic 9. A simplified notation and an algorithm for calculating the SS in ANOVA are presented in Topic 10. Based on the results of the overall ANOVA, the experimenter may conduct a series of post hoc supplemental comparisons to isolate significant main effects and interaction effects as presented in Topic 11.



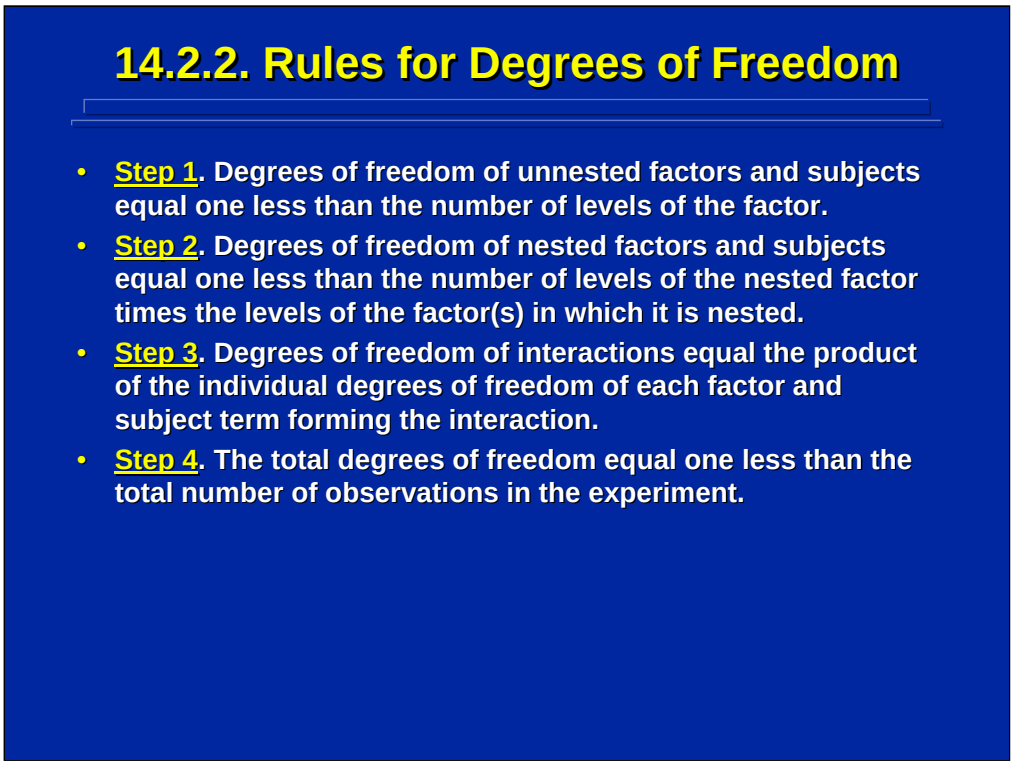

This is a summary of all the algorithms for equal sample size, factorial, ANOVA designs. These are listed in order in which one would use them in calculating a summary table. First state the statistical model. Then calculate the degrees of freedom. Next calculate the sum of squares using the algorithm. Then calculate the mean squares using the algorithm for specifying expected mean squares. Finally go through the steps for determining the F ratios.

Even though computer-based statistical analysis programs are usually used in ANOVA, human factors researchers should always state the sources, degrees of freedom, and error terms for their ANOVA beforehand to avoid conducting an inappropriate ANOVA using the statistical package. These procedural rules facilitate this specification. Details on the uses of these algorithms are presented in Topics 8 and 9 of this reference material.

## 14.2.1. Specification of Statistical Models

- Step 1. Specify an observation as a linear combination of the population mean, main effects, subjects, interactions, and random error where
  - Observation = Y
  - Population Mean = $\mu$
  - Random Error = $\varepsilon$
- Step 2. Specify main effects, subjects, and interactions where
  - Greek letters refer to each factor
  - Subjects = $\gamma$
- Step 3. Denote the levels of each effect by a Roman subscript beginning with letter "i" where
  - Observation, Y, includes all subscripts
  - Levels of each factor have a different subscript
  - Parentheses surround levels of nested effects
  - Random error, $\varepsilon$, is nested in all other effects

This slide summarizes the steps to follow in stating the statistical model of an ANOVA design. Details on using this procedure for specifying the statistical model of an ANOVA design and examples are provided in Topic 9 of this reference material.




This slide summarizes the steps to follow in stating the degrees of freedom of an ANOVA design. Details on using these rules for specifying the df of an ANOVA design and examples are provided in Topic 9 of this reference material.

## 14.2.3. SS Computational Formulae Algorithm

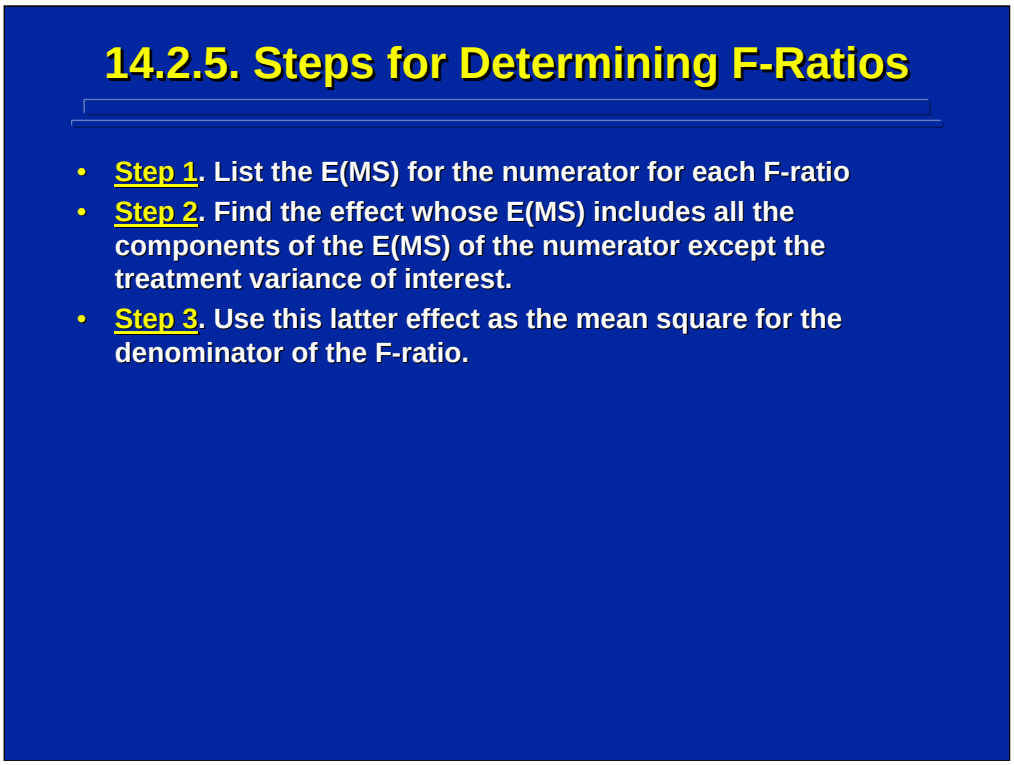
- Step 1. Write the expression for the degrees of freedom of each source of variation and algebraically expand it.
- Step 2. Substitute squared capital letters for each term in the expanded degrees of freedom expression and substitute $T^2$ (the grand total squared) for 1.
- Step 3. Sum all totals across the index(es) of the variable(s) denoted by capital letters, and dot the other index(es). For T merely dot all indexes.
- Step 4. Divide each expression by the number of levels of the dotted index(es).

This slide summarizes the steps to follow in stating the SS computational formulae for an ANOVA design that is based on the simplified notation presented in Topic 10. Details on using this algorithm for calculating the SS of an ANOVA design and examples are provided in Topics 10, 12, and 13 of this reference material.

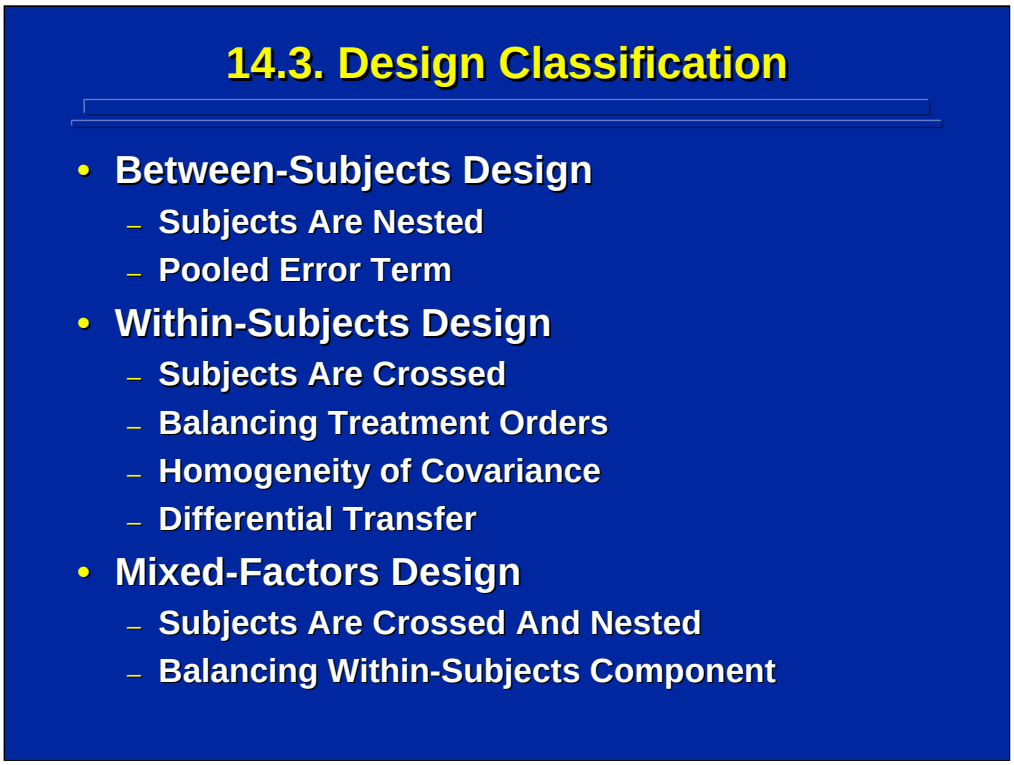
## 14.2.4. Algorithm for Stating E(MS)

- Step 1. Write the appropriate statistical model.
- Step 2. For each random-effect variable, circle the subscript wherever the subscript appears in the model.
- Step 3. To determine the components of the E(MS) for each effect, include:
  - the effect; and
  - other components having the subscript(s) of the effect where all other subscripts are either circled (random variables) or in parentheses (nested within variables).
- Step 4. Begin to list the E(MS) for each effect as a linear combination of the $\sigma^2$ for each component. Note that the subscript for each $\sigma^2$ is the Greek symbol(s) of the component.
- Step 5. To complete the E(MS) listing, multiply each $\sigma^2$ in the resulting linear combination by the number of levels of the factor(s) not involved in defining the component term.

This slide summarizes the steps to follow for stating the E(MS) of an ANOVA design. Details on using this algorithm for specifying the E(MS) of an ANOVA design and examples are provided in Topic 9 of this reference material.




This slide summarizes the steps to follow for determining the F-ratios of an ANOVA design. Details on using this procedure for determining the possible F-ratios of an ANOVA design and examples are provided in Topic 9 of this reference material.




In human factors research involving human subjects, basic ANOVA designs are characterized into three general categories depending on the assignment of subject to treatment conditions. Each category of design has special considerations unique to that type of design. Between-subjects designs, as discussed in Topic 10, are completely randomized designs in which subjects are nested in treatment conditions. The nested subject effect becomes the pooled error term to test all main effects and interactions.

Subjects are crossed with all factors of interest in within-subjects designs as discussed in Topic 12. The main effect of Subjects is removed from the error term which generally results in more sensitive statistical hypothesis testing. Since there are repeated measures of subjects across treatment conditions, balancing the order of treatment presentation must be considered to avoid confounding treatments with order of presentation. The additional assumption of homogeneity of covariance among treatment means and the possibility of differential transfer must be considered in within-subjects designs.

Mixed-factors ANOVA designs are discussed in Topic 13 and have both between-subjects and within-subjects factors. Balancing the within-subjects treatment combinations is necessary to avoid confounding with the presentation order of repeated measures. As in within-subjects designs, a Balanced Latin Square is often used in mixed-factors designs to balance order effects.

## 14.3. Design Classification (Cont'd)

- **Two-Factor, Between-Subjects ANOVA Design**

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_{k(ij)} + \alpha\beta_{ij} + \varepsilon_{l(ijk)}$$

| Source | df | SS | MS | F |
|---|---|---|---|---|
| A | a-1 | $SS_A$ | $MS_A$ | $MS_A/MS_{S/AB}$ |
| B | b-1 | $SS_B$ | $MS_B$ | $MS_B/MS_{S/AB}$ |
| AxB | (a-1)(b-1) | $SS_{AxB}$ | $MS_{AxB}$ | $MS_{AxB}/MS_{S/AB}$ |
| S/AB | ab(n-1) | $SS_{S/AB}$ | $MS_{S/AB}$ | |
| Total | abn-1 | $SS_{total}$ | | |

This slide presents the statistical model and general format for the ANOVA Summary Table for a two-factor, between-subjects design example. Details on the design and the data analysis procedures for between-subjects ANOVA designs are presented in Topics 9 and 10 in this reference material. Note that the $\gamma$ term in the statistical model shows that subjects are nested in both factors A and B, and S/AB is used as the pooled error term for testing the significant main effects of A and B as well as the AxB interaction.

## 14.3. Design Classification (Cont'd)

- **Two-Factor, Within-Subjects ANOVA Design**

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \varepsilon_{l(ijk)}$$

| Source | df | SS | MS | F |
|---|---|---|---|---|
| **Between** | | | | |
| S | n-1 | $SS_S$ | | |
| **Within** | | | | |
| A | a-1 | $SS_A$ | $MS_A$ | $MS_A/MS_{AxS}$ |
| AxS | (a-1)(n-1) | $SS_{AxS}$ | $MS_{AxS}$ | |
| B | b-1 | $SS_B$ | $MS_B$ | $MS_B/MS_{BxS}$ |
| BxS | (b-1)(n-1) | $SS_{BxS}$ | $MS_{BxS}$ | |
| AxB | (a-1)(b-1) | $SS_{AxB}$ | $MS_{AxB}$ | $MS_{AxB}/MS_{AxBxS}$ |
| AxBxS | (a-1)(b-1)(n-1) | $SS_{AxBxS}$ | $MS_{AxBxS}$ | |
| **Total** | abn-1 | $SS_{total}$ | | |

This slide presents the statistical model and general format for the ANOVA Summary Table for a two-factor, within-subjects design example. Details on the design and the data analysis procedures for within-subjects ANOVA designs are presented in Topics 9 and 12 in this reference material. Note that the $\gamma$ term in the statistical model shows that subjects are crossed with both factors A and B. The main effect of subjects, S, is removed from the error term as a between-subjects component, and the interactions of S with A, B, and AxB are used to test those effects, respectively, as within-subjects components.

Balancing of presentation order needs to be considered for all the $ab_{ij}$ treatment combinations in this repeated measures design example. The choice of complete counterbalancing, a partially Balanced Latin Square, or random assignment of treatment orders across subjects has implications for the number of subjects, n, used in the experiment.

## 14.3. Design Classification (Cont'd)

- **Two-Factor, Mixed-Factors ANOVA Design**

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_{k(i)} + \alpha\beta_{ij} + \beta\gamma_{jk(i)} + \varepsilon_{l(ijk)}$$

| Source | df | SS | MS | F |
|---|---|---|---|---|
| **Between** | | | | |
| A | a-1 | $SS_A$ | $MS_A$ | $MS_A/MS_{S/A}$ |
| S/A | a(n-1) | $SS_{S/A}$ | $MS_{S/A}$ | |
| **Within** | | | | |
| B | b-1 | $SS_B$ | $MS_B$ | $MS_B/MS_{BxS/A}$ |
| BxA | (a-1)(b-1) | $SS_{BxA}$ | $MS_{BxA}$ | $MS_{BxA}/MS_{BxS/A}$ |
| BxS/A | a(b-1)(n-1) | $SS_{BxS/A}$ | $MS_{BxS/A}$ | |
| **Total** | abn-1 | $SS_{total}$ | | |

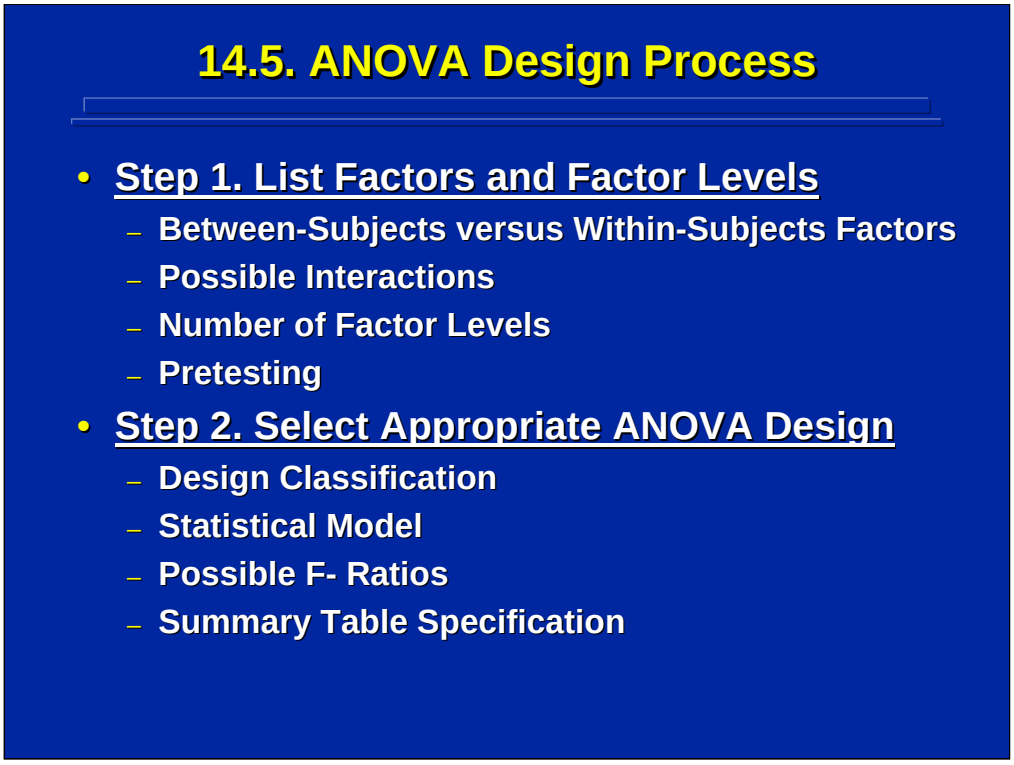
This slide presents the statistical model and general format for the ANOVA Summary Table for a two-way, mixed-factors design example. Details on the design and the data analysis procedures for mixed-factors ANOVA designs are presented in Topics 9 and 13 in this reference material. Note that the $\gamma$ term in the statistical model shows that subjects are nested in Factor A and crossed with Factor B. Factor A is tested with S/A error term as a between-subjects component. Both factor B and the AxB interaction are tested by the BxS/A interaction error term as within-subjects components.

Balancing of presentation order needs to be considered for all the levels of Factor B in this repeated measures design example. The choice of complete counterbalancing, a partially Balanced Latin Square, or random assignment of treatment orders across subjects has implications for the number of subjects, n, used in the experiment just as in the previous within-subjects design example.

## 14.4. n-Factor Design Generalizations

- **n-Factor ANOVA Design Generalizations**
  - Can include any number of fixed-effects factors of interest.
  - All rules, procedures, and algorithms apply.
  - All factors of interest are crossed and can interact.
  - The subject effect is crossed with all within-subjects factors of interest and can interact with them.
  - Subjects are random-effects, and sample size is equal.
  - Assuming factors of interest are fixed-effects,
    - Subjects are nested within all between-subjects factors of interest, and this subject effect is the error term for all between-subjects F-tests.
    - The interaction of the within-subjects effect with the subject effect is the error term for all F tests on the within-subjects effect as well as its interactions with the between-subjects effects.

Generalizations can be stated for basic ANOVA designs and analyses used in human factors and ergonomics research regardless of the number of factors included in the experiment. This slide summarizes generalizations for any n-factor design. Specific generalizations for between-subjects, within-subjects, and mixed-factors are presented separately in Topics 10, 12, and 13, respectively, in this reference material.




The experimenter needs to consider the overall process in choosing an ANOVA design rather than just the fundamental mechanics of ANOVA design and analysis. A six-step process is presented on this slide and the next two slides. This process begins by listing the factors of interest to the research in Step 1. These factors need to be defined as either between-subjects factors or within-subjects factors. If factors are included in the same factorial ANOVA, they can possibly interact with each other. The various levels of each factor need to be specified, and this process often requires pretesting before making a final decision on the factors and factor levels to be investigated in a reasonably sized factorial ANOVA design.

Once the factors and factor levels of interest are selected, the experimenter selects the appropriate ANOVA design in Step 2. The design is specified as a between-subjects, within-subjects, or mixed-factors ANOVA design. A statistical model should be specified for this design based on the crossing or nesting of factors with subjects. This statistical model specifies all the effects that can be estimated from the data collected in the experiment. The experimenter should determine if all the effects of interest can be tested for significance by listing all the possible F-tests in the design. Finally, the experimenter should list an ANOVA Summary Table that includes at least the Sources, df, and F-tests in general terms to determine if the design is adequate.

## 14.5. ANOVA Design Process (Cont'd)

- **Step 3. Determine Appropriate Sample Size, n**
  - Subject Availability
  - Resulting Treatment Conditions
  - Balancing Practice Effects
- **Step 4. Establish Data Collection Procedure**
  - Instructions
  - Practice
  - Data Recording
- **Step 5. Develop Data Analysis Plan**
  - Primary Analysis
  - Secondary Analysis

Once a candidate factorial design is chosen, the experimenter determines the appropriate sample size, n, for the experiment in Step 3. A large number of between-subjects factors requires many subjects, and subject availability needs to be assessed. If within-subjects factors are included in the design, they define the number of treatments each subject receives and has implications for how long each subject needs to participate. In addition, a procedure for balancing presentation order must be chosen for the repeated measure. Complete counterbalancing and Balanced Latin Square procedures have implication for minimum sample size as discussed in Topic 12.

Actual data collection occurs in Step 4. Care must be taken to provide adequate instructions and practice on the experimental task. The researcher should provide a procedure that insures proper data recording for subsequent statistical analysis.

The plan for actual statistical analysis is developed in Step 5. Primary ANOVA procedures are discussed in Topics 10, 12, and 13 to provide overall F-tests of main effects and interactions in between-subjects, within-subjects, and mixed-factors designs, respectively. Subsequent post hoc tests used to isolate significant main effects and interactions are described in Topic 11.

## 14.5. ANOVA Design Process (Cont'd)

- **Step 6. Interpret ANOVA Results**
  - Graphing Techniques
  - Main Effects
    - Post Hoc Comparisons
  - Interactions
    - Simple Effects
    - Trend Analyses
    - Post Hoc Comparisons
  - Supplemental Data Analysis
  - Verbal Description

Interpretation of the results of an experiment is the key concern of the human factors researcher. Statistical analysis of basic ANOVA designs is merely a tool to aid in interpretation. The experimenter should always graph the results of main effects and interactions to assist in possible interpretations of significant effects. But, additional post hoc analyses are needed to verify these interpretations if more than two levels of any factor are observed in the experiment. Various post hoc analysis comparison techniques are covered in Topic 11. Simple effects tests, trend analyses, and post hoc comparisons to isolate interaction effects are also covered in Topic 11 in this reference material.

The experimenter often draws upon supplemental data analysis to facilitate interpretation of the significant main effects and interactions found in the basic ANOVA design. Collection procedures and nonparametric data analysis procedures for these supplemental data are presented in Section 2 of this reference material.

Once all the primary, post hoc, and supplementary data analyses are completed, the experimenter can provide a complete verbal description of the results. Precise and succinct presentations that provide clear descriptions of results are needed for successful communication to the scientific community at large.

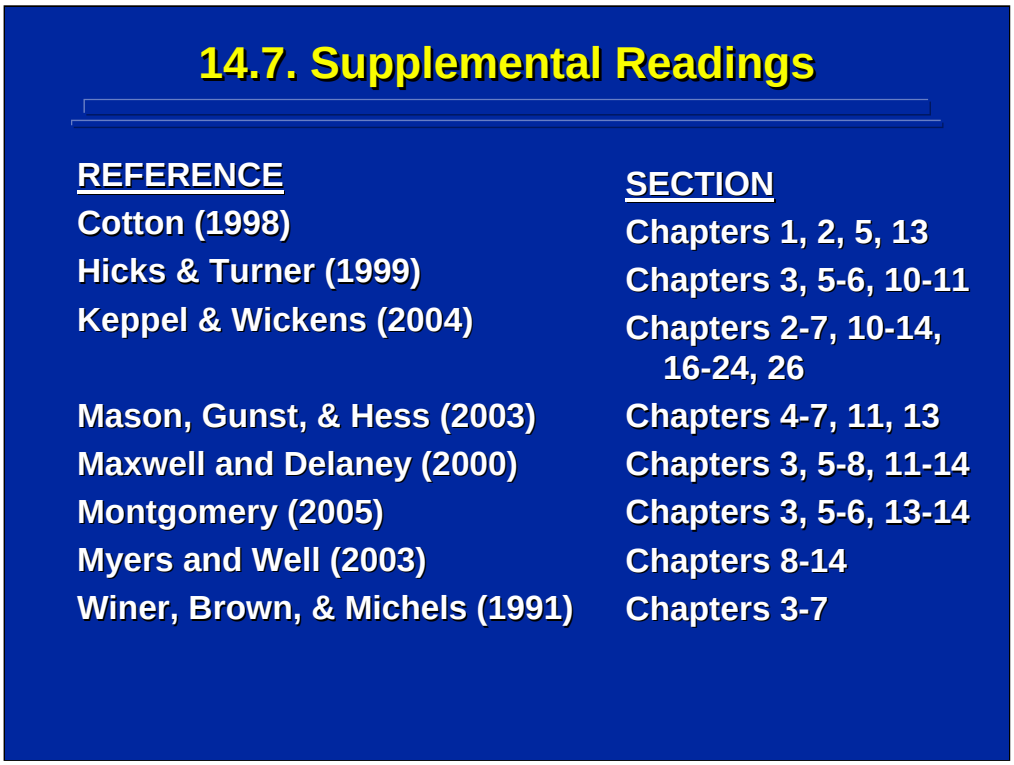
## 14.6. Summary

- General Approach to Basic ANOVA
  - Terminology
  - Simplified Notation
- Experimental Design Process for Basic ANOVA
  - Design Classification and Statistical Model
  - Design Considerations
- Data Analysis for Basic ANOVA
  - ANOVA Assumptions
  - Computational Procedures, Rules, and Algorithms
  - ANOVA Summary Table
  - Post Hoc Analysis of Main Effects and Interactions

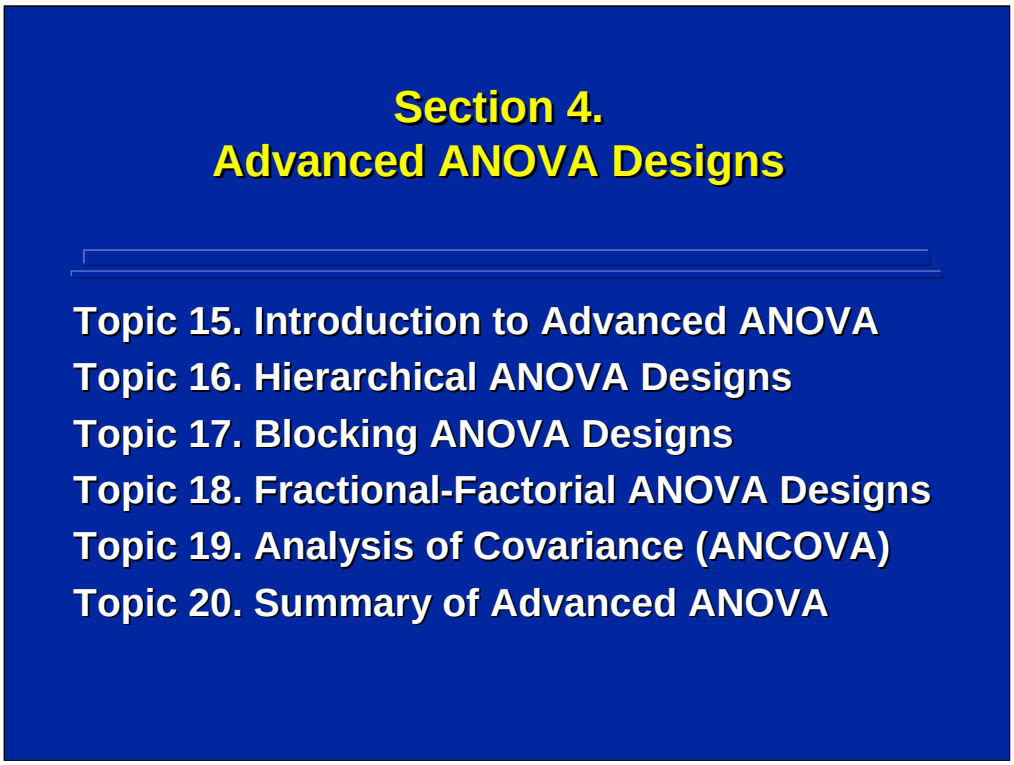
By way of a summary for this section, this reference material provides a general approach to experimental design and analysis of basic ANOVA that can be used by human factors and ergonomics researchers. The approach to ANOVA is described using a fundamental terminology and simplified design notation throughout this section.

Basic procedures, rules, and algorithms exist for generating any ANOVA design and conducting subsequent data analysis. In human factors research, the experimenter has a choice of using between-subjects, within-subjects, or mixed-factors designs. The advantages and disadvantages of design alternative and the nature of the real world variable must be considered in choosing one of these three design alternatives.

Once a design is chosen, appropriate ANOVA assumptions for that design category need to be considered. Interpretation of results is the key to any successful experiment. Primary analyses, post hoc analyses, supplemental analyses, and data graphing are all important to making clear interpretations of results.




A summary of chapters dealing with basic ANOVA topics in common experimental design textbooks used by human factors researchers are listed on this slide. The chapters in Keppel and Wickens (2004) and Winer et al. (1991) most closely follow the procedures covered in this topic.




Section 4 covers a variety of techniques that build upon basic ANOVA and allow the experimenter to investigate special circumstances in human factors and ergonomics research. This section covers the following topics:

Topic 15 – introduction to advanced ANOVA designs, quasi-F ratios, and randomized blocks designs;

Topic 16 – partial and complete hierarchical ANOVA designs;

Topic 17 – simple and complex blocking ANOVA designs;

Topic 18 – fractional-factorial ANOVA designs and Latin square designs;

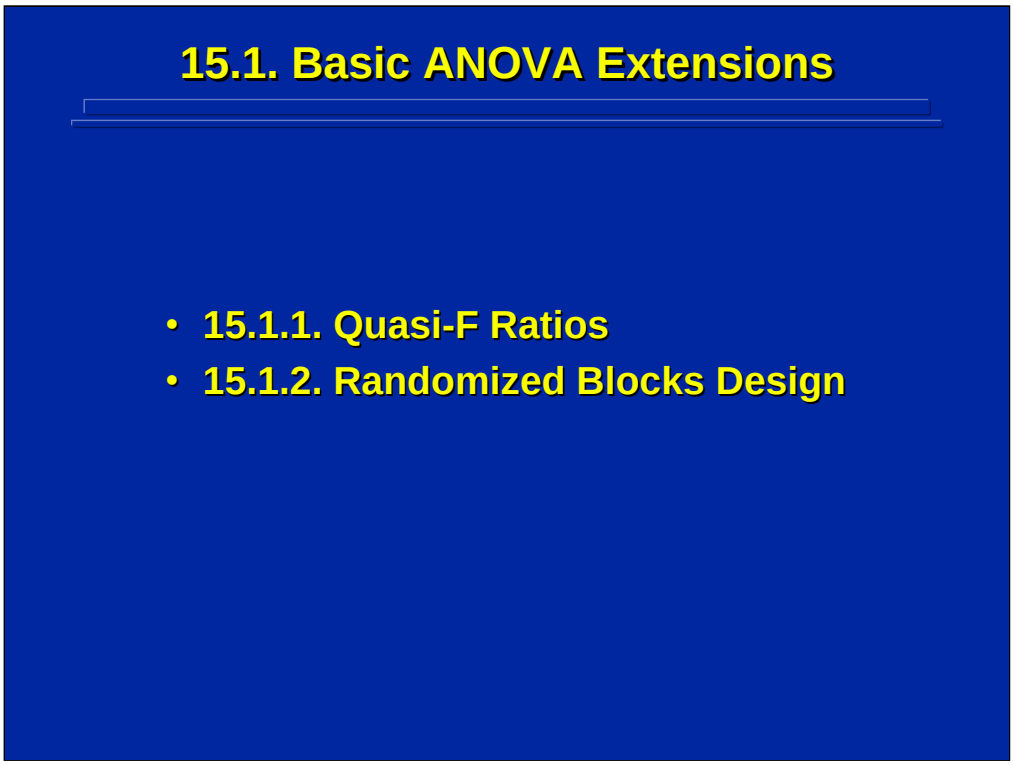
Topic 19 – review of correlation, simple regression, and analysis of covariance; and

Topic 20 – summary of advanced ANOVA designs.

# Topic 15. Introduction to Advanced ANOVA

15.1. Basic ANOVA Extensions
15.1.1. Quasi-F Ratios
15.1.2. Randomized Blocks Design
15.2. Advanced ANOVA Design and Analysis
15.3. Summary
15.4. Supplemental Readings

Advanced ANOVA design topics use the basic ANOVA procedures, rules, and algorithms described in Section 3 of this reference material. Two examples of extending basic ANOVA (i.e., quasi-F ratios and randomized blocks designs) are presented in detail. In addition, this introduction provides an overview of several special purpose ANOVA design and analysis procedures that satisfy various constraints present in human factors and ergonomics research. Detailed discussions of these procedures are presented in subsequent topics in Section 4.




Two extensions to basic ANOVA are discussed in this topic. First, quasi-F procedures are described to estimate legitimate F-ratios when random-effects variables of interest are considered in ANOVA experimental designs. Second, randomized blocks ANOVA designs are described as a means of increasing the sensitivity of between-subjects ANOVA designs by refining the error term.

## 15.1.1. Quasi- F Ratios

- Error Terms of F-Ratios
- Random Effects Variable Example

- E(MS) for Two-Factor, Within-Subjects Design (A, B, and S Random)

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \varepsilon_{l(ijk)}$$

$$E(MS_A) = bn\sigma_\alpha^2 + n\sigma_{\alpha\beta}^2 + b\sigma_{\alpha\gamma}^2 + \sigma_{\alpha\beta\gamma}^2 + \sigma_\varepsilon^2$$
$$E(MS_B) = an\sigma_\beta^2 + n\sigma_{\alpha\beta}^2 + a\sigma_{\beta\gamma}^2 + \sigma_{\alpha\beta\gamma}^2 + \sigma_\varepsilon^2$$
$$E(MS_S) = ab\sigma_\gamma^2 + b\sigma_{\alpha\gamma}^2 + a\sigma_{\beta\gamma}^2 + \sigma_{\alpha\beta\gamma}^2 + \sigma_\varepsilon^2$$
$$E(MS_{AxB}) = n\sigma_{\alpha\beta}^2 + \sigma_{\alpha\beta\gamma}^2 + \sigma_\varepsilon^2$$
$$E(MS_{AxS}) = b\sigma_{\alpha\gamma}^2 + \sigma_{\alpha\beta\gamma}^2 + \sigma_\varepsilon^2$$
$$E(MS_{BxS}) = a\sigma_{\beta\gamma}^2 + \sigma_{\alpha\beta\gamma}^2 + \sigma_\varepsilon^2$$
$$E(MS_{AxBxS}) = \sigma_{\alpha\beta\gamma}^2 + \sigma_\varepsilon^2$$

- F Ratios

$$F_A = (MS_A)/? \qquad F_B = (MS_B)/? \qquad F_{AxB} = (MS_{AxB})/(MS_{AxBxS})$$

Quasi-F ratios are used when no error term exists to test an effect of interest in an ANOVA design. Subsequently, an error term is constructed by combining effects to estimate a legitimate F-ratio. This situation can occur when effects of interest in an ANOVA design are considered random-effects rather than fixed-effects variables.

Consider the two-factor, within-subjects design shown on this slide. Factors A, B and subjects are all random-effects variables. The E(MS) for this design can be generate by the algorithm described in Topic 9 and shown in the center of the slide. Note that the only legitimate F test in this design is the AxB interaction based on the E(MS), and no error term exists for either the A or B main effects. Quasi-F ratios can be constructed to test the significance of each of the two main effects by combining mean squares in the error term.

Two steps are needed in quasi-F tests to estimate the standard F test. Quasi-F tests require both the construction of the observed F-ratio and the Satterthwaite (1946) correction for the degrees of freedom used in determining the tabled value of F.

## 15.1.1. Quasi-F Ratios (Cont'd)

- **Constructing Quasi-F Test of the Main Effect of A**
  - **Numerator**

$$E(MS_A) = bn\sigma_\alpha^2 + n\sigma_{\alpha\beta}^2 + b\sigma_{\alpha\gamma}^2 + \sigma_{\alpha\beta\gamma}^2 + \sigma_\varepsilon^2$$

  - **Denominator**

$$E(MS_{denominator}) = n\sigma_{\alpha\beta}^2 + b\sigma_{\alpha\gamma}^2 + \sigma_{\alpha\beta\gamma}^2 + \sigma_\varepsilon^2$$

$$[n\sigma_{\alpha\beta}^2 + \sigma_{\alpha\beta\gamma}^2 + \sigma_\varepsilon^2] + [b\sigma_{\alpha\gamma}^2 + \sigma_{\alpha\beta\gamma}^2 + \sigma_\varepsilon^2] - [\sigma_{\alpha\beta\gamma}^2 + \sigma_\varepsilon^2]$$

$$E(MS_{denominator}) = E(MS_{AxB}) + E(MS_{AxS}) - E(MS_{AxBxS})$$

  - **Quasi-F Ratio**

$$F'_A = (MS_A) / [(MS_{AxB}) + (MS_{AxS}) - (MS_{AxBxS})]$$

$$F''_A = [(MS_A) + (MS_{AxBxS})] / [(MS_{AxB}) + (MS_{AxS})]$$

This slide shows the quasi-F ratio for testing Factor A. It is constructed by combining various E(MS) in the denominator in order to obtain only the A effect, $bn\sigma_\alpha^2$, in the numerator of the F ratio with the other effects cancelled out by the denominator according to the E(MS) of the two-factor, random-effects design. The resulting Quasi-F ratio, $F_A'$, has three mean squares (i.e., $MS_{AxB}$ + $MS_{AxS}$ + $MS_{AxBxS}$) in the denominator rather than just one mean square in basic, fixed-effects ANOVA designs. Note that one of the three mean squares, $MS_{AxBxS}$, is subtracted from the other two, which allows the opportunity of obtaining a negative total mean square value in the denominator. To avoid this possibility, $MS_{AxBxS}$ is added to the numerator to form a quasi-F ratio referred to as $F_A''$. Usually a F″ is used instead of a F′ when constructing quasi-F ratios to avoid a negative F value even though there is a slight risk of an inflated $\alpha$ error resulting from the higher observed F-ratio in F".

## 15.1.1. Quasi-F Ratios (Cont'd)

- **Satterthwaite (1946) $F_{Tabled}$ Value Correction**

  - **Correction for $df_{tabled}$ Values**

    $F' = (a) / (c + d - b)$

    $df_{numerator} = df_a$

    $df_{denominator} = (c + d - b)^2 / [(c^2/df_c) + (d^2/df_d) + (b^2/df_b)]$

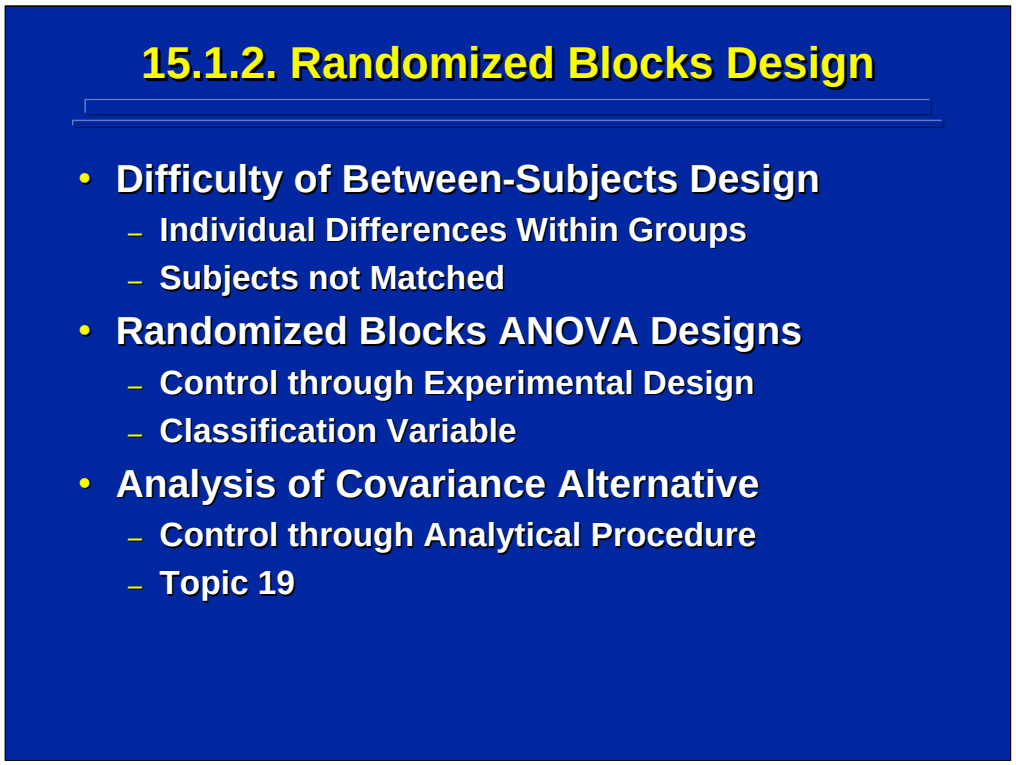
    $F'' = (a + b) / (c + d)$

    $df_{numerator} = (a + b)^2 / [(a^2/df_a) + (b^2/df_b)]$

    $df_{denominator} = (c + d)^2 / [(c^2/df_c) + (d^2/df_d)]$

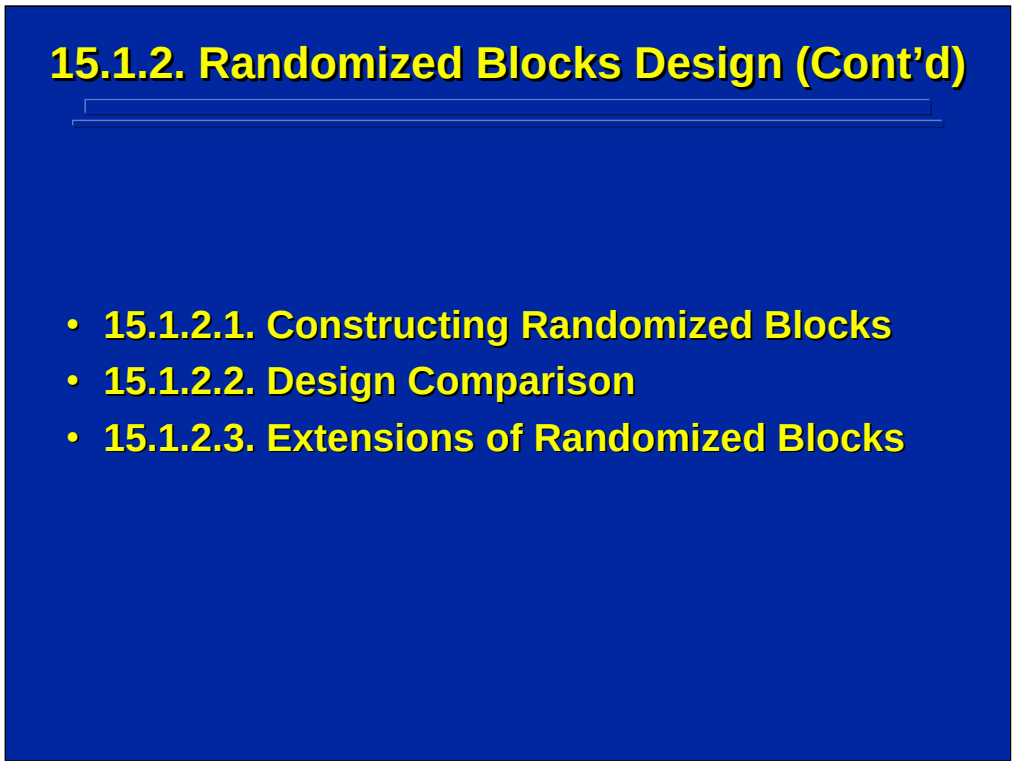
Satterthwaite (1946) provided a correction for the tabled value of F sampling distributed when quasi-F ratios are used to estimate the standard F ratio. The correction is determined by changing the degrees of freedom for the numerator and denominator in the standard F table according to the formulae given on this slide for both F′ and F″.

Note that "a", "b", "c", and "d" given in these formulae refer to the value of the various MS components of F′ and F″. Specifically, a = $MS_A$, b = $MS_{AxB}$, c = $MS_{AxS}$, and d = $MS_{AxBxS}$ in the example described on the previous slide.

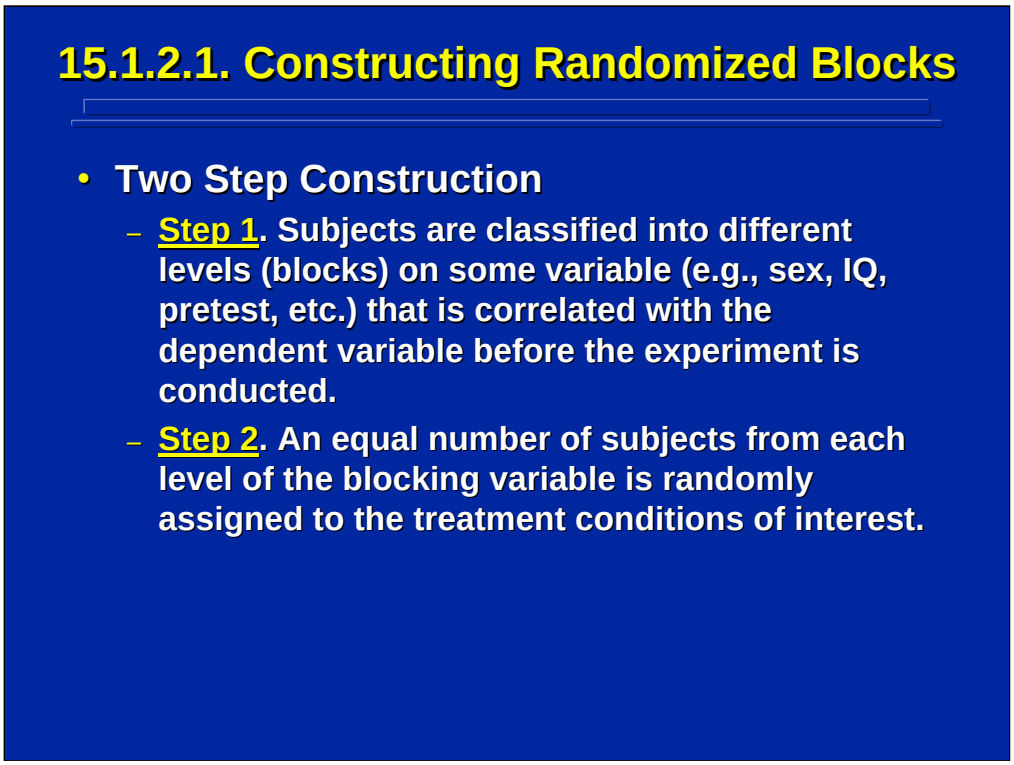



The major difficulty in using a between-subjects design in human factors and ergonomics research is that a pooled error term is used resulting in an insensitive F test as compared to an alternative within-subjects design. This pooled error term combines the individual differences of subjects within a group. Since subjects are not generally matched across groups these individual differences are often one of the largest sources of variation in a human factors experiment.

To minimize individual difference effects, subjects can be categorized beforehand into different levels, or blocks, on a classification variable that is known to correlate with the dependent variable. Classification variables such as gender, experience, and aptitude are often used in human factors research. Subsequently, an equal number of subjects in each classification (i.e., block) is randomly assigned to each treatment condition in the design, hence forming a “randomized blocks” experimental design. The block effect is removed from the error term in the between-subjects ANOVA thereby making the F test on the between-subjects factor more sensitive.

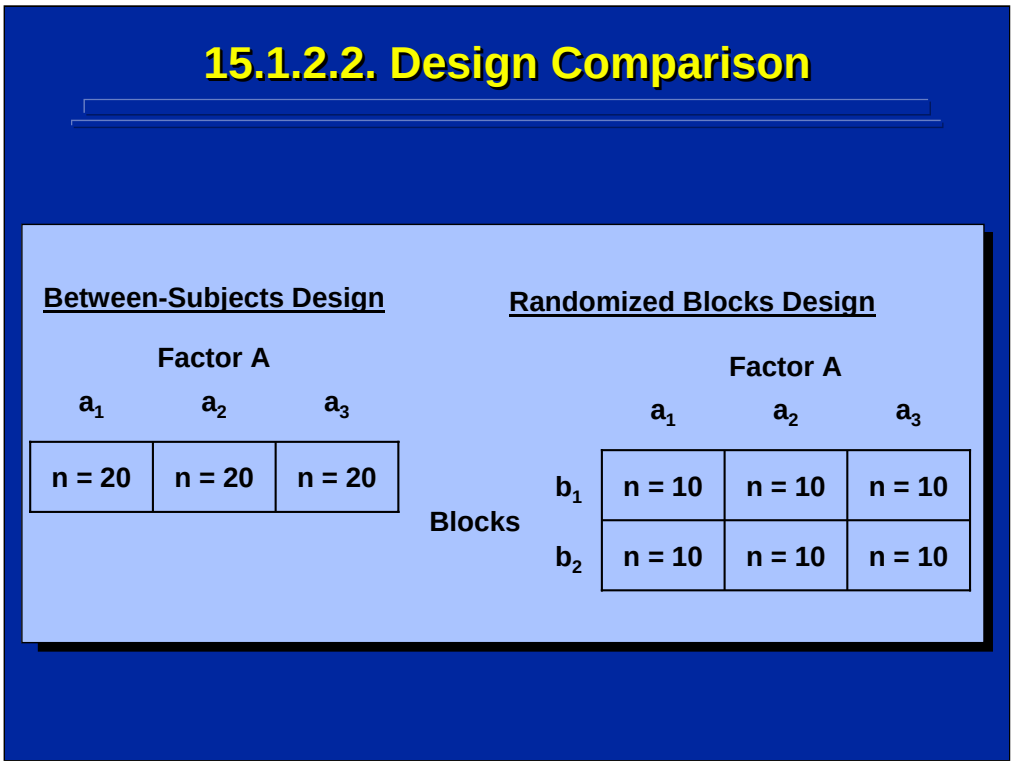



First, a two step procedure for generating a randomized blocks design is described. Second, a comparison between the randomized block ANOVA design and its between-subjects ANOVA design counterpart is presented to demonstrate the differences in sensitivity between the two ANOVA design alternatives. Finally, possible extensions of randomized block designs are discussed.




Every randomized blocks design is constructed by a two step procedure. First, subjects are classified into blocks before they are assigned to a treatment condition in the between-subjects design. The classification variable must be significantly correlated with the dependent variable in order for the blocking to be effective. Often the classification value is known (e.g., gender) or exists in records (e.g., educational background). If classification is not known beforehand, subjects need to be pretested on the classification variable (e.g., IQ test, verbal aptitude, spatial aptitude, etc.) before assignment. This pre-testing requires additional time and cost in conducting the experiment.

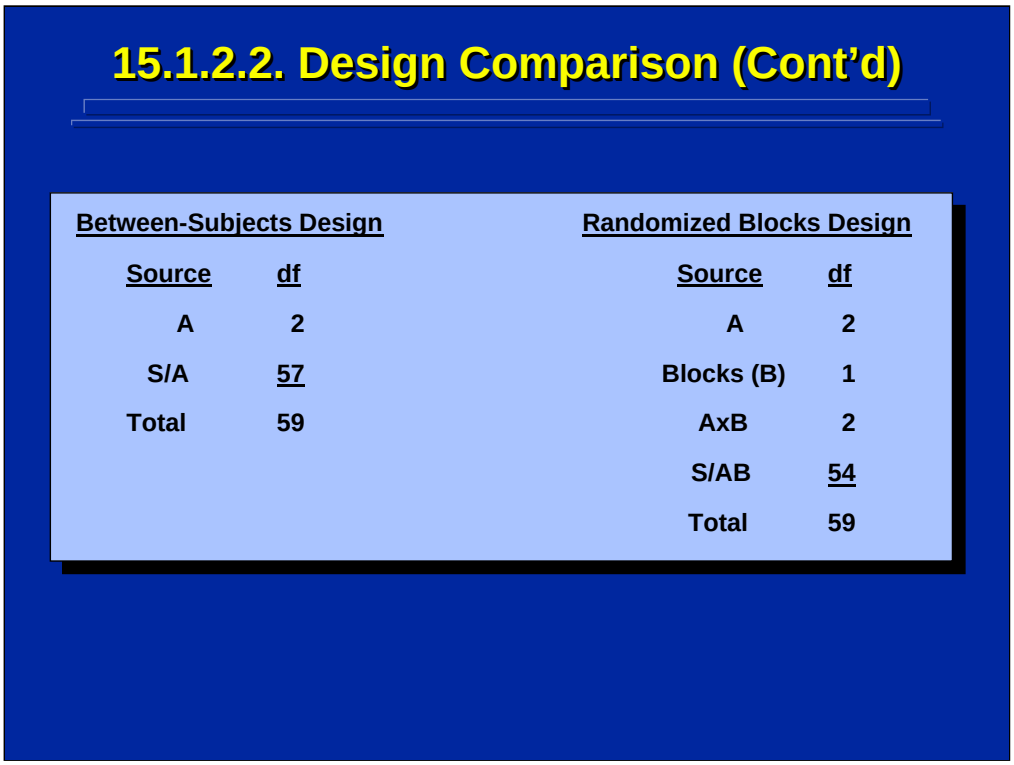
Second, an equal number of subjects in each level of the classification variable is randomly assigned to each cell in the between-subjects design. Often it may be necessary to pretest more subjects than the minimum number of subjects needed to obtain equal sample size, because the subjects volunteering for the experiment usually are not equally represented at each level of the blocking variable.




This slide compares a schematic of a three-level, one-factor, between-subjects design to a schematic of its counterpart randomized blocks design that has a blocking variable (e.g., gender) with two levels (males and females). Note both between-subjects design alternatives require a total of 60 subjects in the experiment.

In the between-subjects design, 20 subjects are randomly assigned to each of the three levels of Factor A. In the randomized blocks counterpart, 10 males and 10 females are assigned to each of the three levels of Factor A to yield a total of 20 different subjects in each level of Factor A. Consequently, 30 males and 30 females are needed for random assignment in the randomized block design, but gender does not need to be equally represented in the between-subject design counterpart. Consequently, subject recruitment is less complicated in the between-subjects design.

Note that the randomized block design essentially just adds a second factor, Blocks, to the between-subjects design. The effect of the blocking variable is usually of no research interest to the experiment and is used only as a way of classifying subjects to control for individual differences in order to make the between-subjects design more sensitive. Hence, blocking variables are usually categorical variables not manipulated variables of experimental interest.




This slide compares the Sources and degrees of freedom (df) of the one-factor, between-subjects design with its randomized blocks design counterpart. Both designs have 59 total df, and all ANOVA calculations for the main effect of Factor A (2 df) are the same. The Blocks (1 df) main effect and the Factor A by Blocks (2 df) interaction are removed from the error term in the randomized blocks design. The df for these two effects are subtracted from the pooled error term. Calculations for Blocks and the Factor A by Blocks interaction follow the standard rules, procedures, and algorithms presented in Section 3.

Note that the error term for Factor A (2 df) is S/A (57 df) in the between-subjects design and is S/AB (54 df) in the randomized blocks design. As long as the Blocking variable is significantly correlated with the dependent variable, the F-test for Factor A is more sensitive in the randomized blocks design than in the between-subjects design even though it has fewer df in the denominator since a significant amount of variability is removed from the error term.

## 15.1.2.3. Extensions of Randomized Blocks

- Optimal Number of Blocks (Feldt, 1958)
- Alternative Blocking Designs
  - Higher-Order, Between-Subjects Designs
  - More Than One Blocking Variable
  - Post-Hoc Blocking vs. Analysis of Covariance
- Conclusions
  - Increase in Precision
  - Increase in Time and Cost
    - Determine Correlation
    - Pre-test Subjects
    - Number of Subjects

Randomized blocks designs can be extended in several ways. Guidelines for choosing the optimal number of blocks are presented in a table by Feldt (1958) and depend on the degree of correlation between the blocking variable and the dependent variable, the number of treatment levels of interest, and the total number of available subjects for the experiment. Randomized blocks designs can be easily extended to higher-order factorial designs and more than one blocking variable can be included. A blocking variable can be added post hoc after completion of data collection, but it is unlikely that the experimenter can maintain equal sample size thereby reducing sensitivity. Consequently, post-hoc blocking is usually not used and analysis of covariance as described in Topic 19 is used instead.

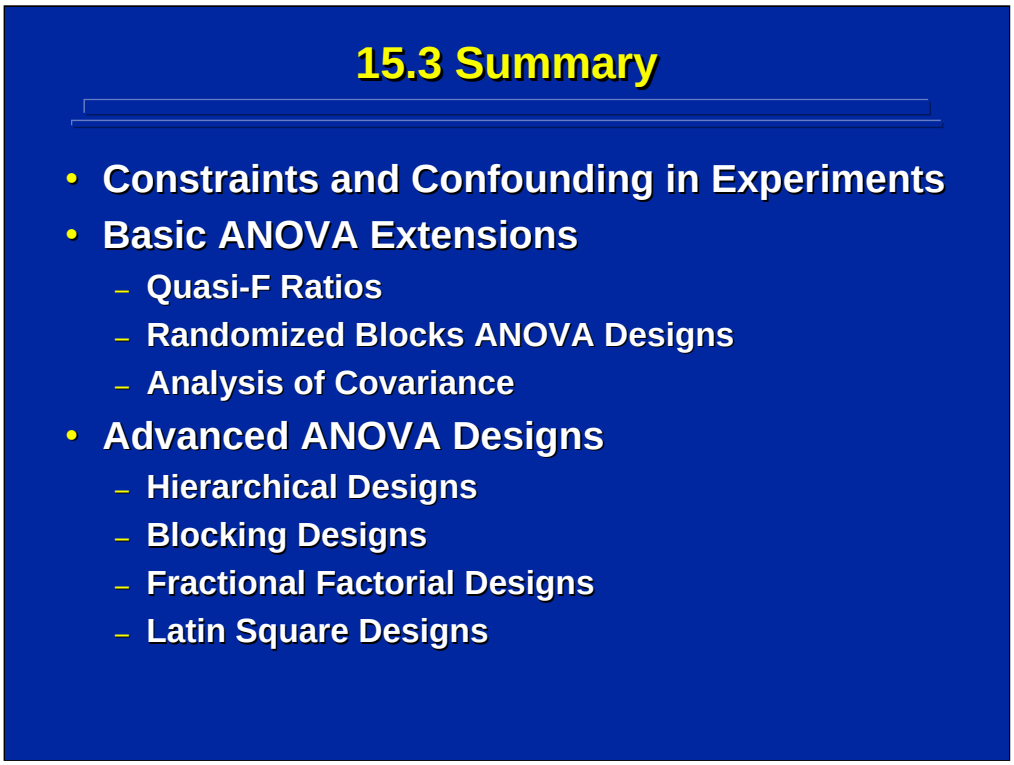
The pros and cons of using randomized blocks designs must be considered carefully. If a known blocking variable exists in the literature, it can be used effectively to increase the precision of the between-subjects design. But, blocking requires additional effort in determining an appropriate blocking variable that correlates with the dependent variable, possibly by pretesting the subjects, and by probably needing to recruit more than the minimum number of subjects in order to maintain equal sample size.

## 15.2. Advanced ANOVA Design and Analysis

- ANOVA Design Constraints
  - Nested Factors Of Interest
  - Control Of Nuisance Variables
  - Limited Data Collection
- Advanced ANOVA Design Alternatives
  - Hierarchical Designs – Topic 16
  - Blocking Designs – Topic 17
  - Fractional-Factorial Designs – Topic 18
  - Latin Square Designs – Topic 18
- Regression Analysis in Experimentation
  - Review of Correlation and Simple Regression
  - Analysis of Covariance (ANCOVA) – Topic 19

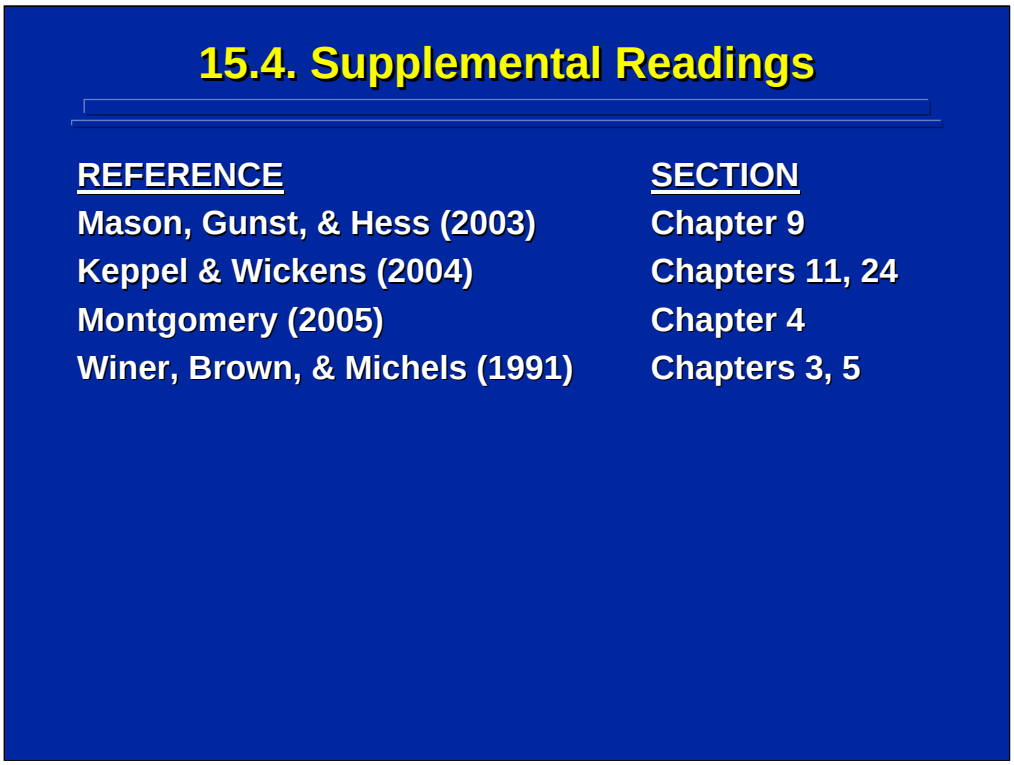
The remaining topics in Section 4 are devoted to other advanced ANOVA design and analysis procedures. Advanced ANOVA designs address various experimental design constraints. Topic 16 covers hierarchical designs that allow investigation of factors of interest that are nested. Blocking designs that control for nuisance variables that could confound the results of the experiment are discussed in Topic 17. Fractional-factorial designs and Latin square designs that allow the experimenter to conduct experiments when a complete higher-order factorial ANOVA design cannot be used are described in Topic 18.

Topic 19 introduces the use of regression analysis by reviewing correlation and simple regression. An application of simple regression is discussed in terms of analysis of covariance (ANCOVA) that can be used as an alternative to using randomized blocks in between-subjects ANOVA designs as discussed in this topic. Further applications of regression analysis are described in Section 5 in terms of building empirical models through experimentation.




By way of summary, this introductory topic on advanced ANOVA addresses some real-world constraints that must be considered in experimental design. Two basic ANOVA extensions are covered in detail. First, the experimenter may need to construct quasi-F ratios in the ANOVA when variables of interest exist in the real world as random-effect factors and no legitimate error term exists. Second, the experimenter may increase the sensitivity of a between-subjects ANOVA design by removing subject variability from the error term if subjects can be categorized by a factor known to be correlated with the dependent variable. In such circumstances a randomized block ANOVA design is appropriate. Alternatively, an analysis of covariance as described in Topic 19 can be considered.

Other experimental design constraints are described in additional advanced ANOVA topics in Section 4 dealing with hierarchical designs, simple and compound blocking designs, $2^k$ fractional-factorial designs, and Latin square designs. These advanced ANOVA designs handle situations in which the factors of interest are nested, multiple sessions or multiple experimenters are required for data collection, and only a fraction of the full factorial design can be investigated due to budget and time constraints.

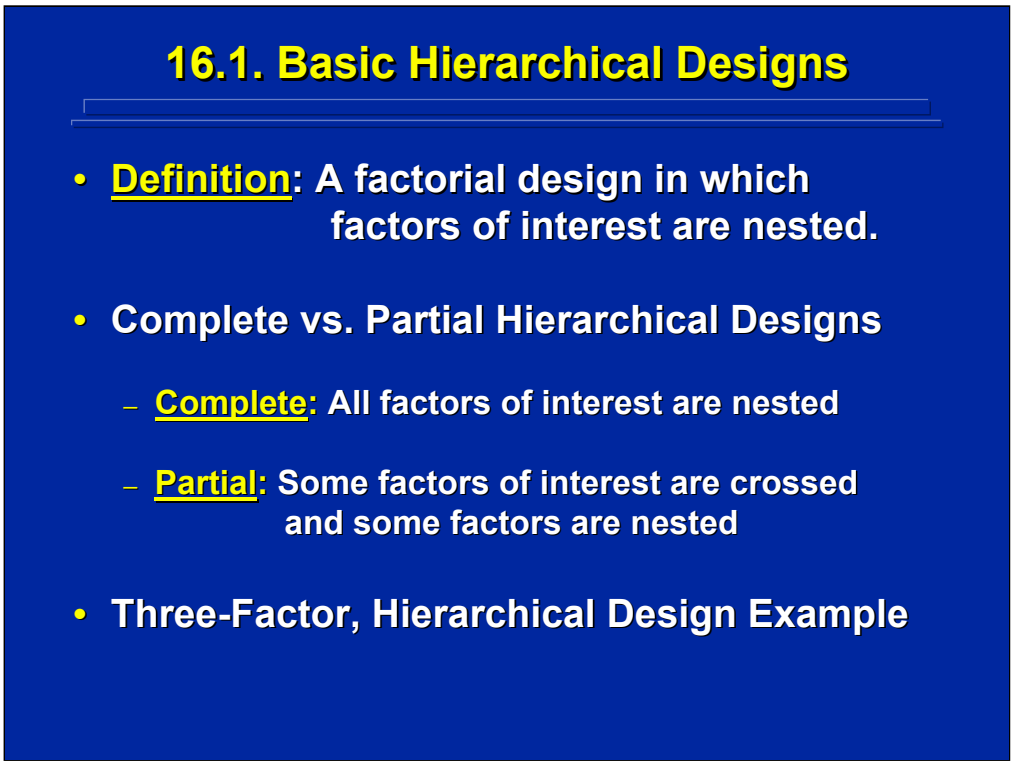



The chapters in the texts listed on this slide discuss quasi-F ratios and randomized block designs.

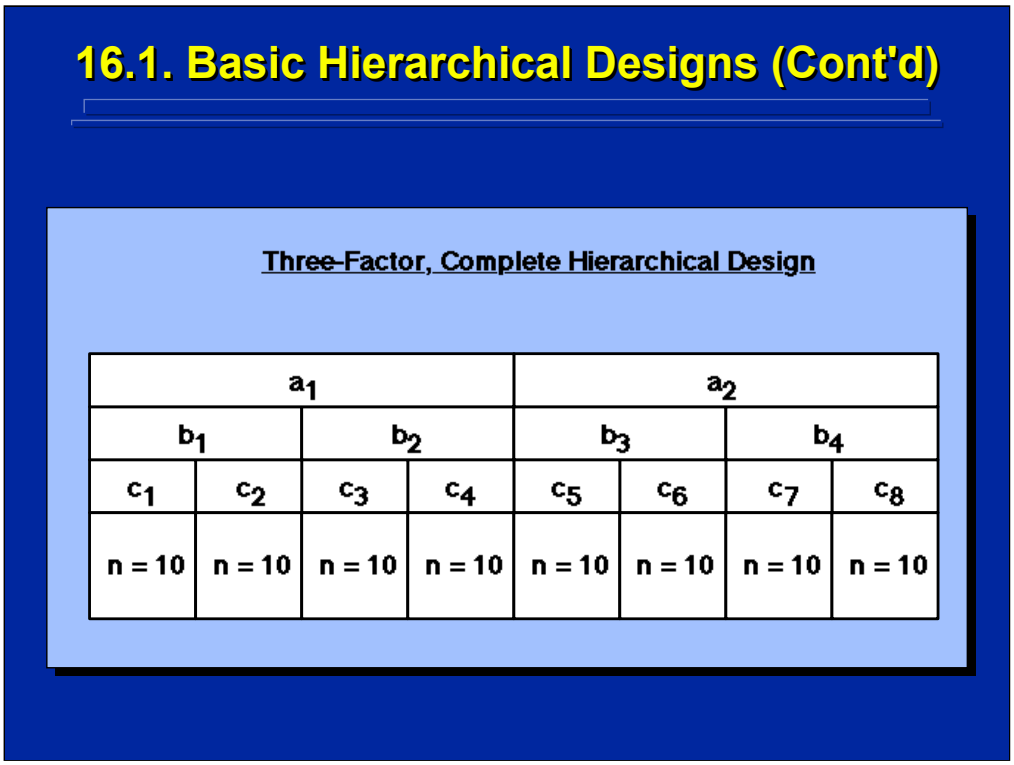
# Topic 16. Hierarchical ANOVA Designs

16.1. Basic Hierarchical Designs
16.1.1. Between-Subjects Designs
16.1.2. Within-Subjects Designs
16.1.3. Mixed-Factors Designs
16.2. Hierarchical Design Examples
16.2.1. Complete Hierarchical Design
16.2.2. Partial Hierarchical Design
16.3. Summary
16.4. Supplemental Readings

Hierarchical designs are ANOVA designs that include nested factors of interest. This topic describes both the basic layout and the ANOVA computations involved in this class of experimental designs. These basic procedures can be generalized to higher-order hierarchical designs. This topic ends with a general summary of the considerations of using hierarchical designs in human factors and ergonomics research. Suggested readings on hierarchical designs in standard experimental design texts are also provided.

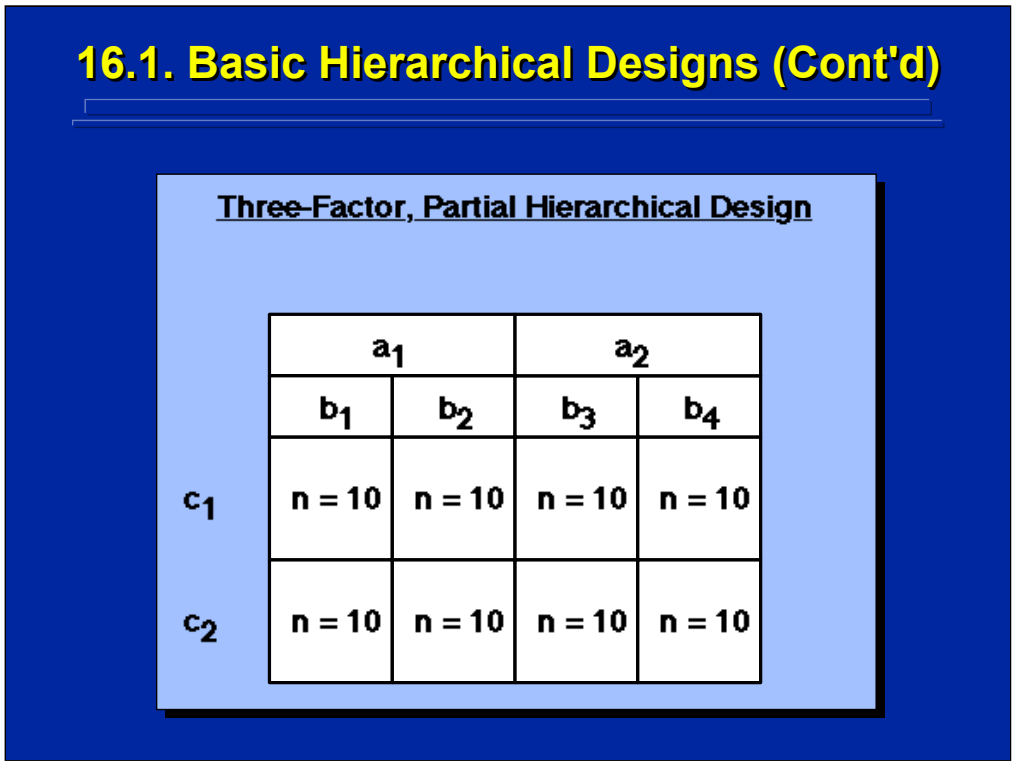



Basic ANOVA designs cover situations in which all the factors of interest are crossed and only subjects are nested within factors. At times, some of the factors of interest may be nested thereby forming a hierarchical design. When all factors in the experimental design are nested, the design is referred to as a complete hierarchical design. When some factors are crossed and some are nested, the design is called a partial hierarchical design. A common three-factor design is used throughout this topic to distinguish the various hierarchical design alternatives and computational procedures.

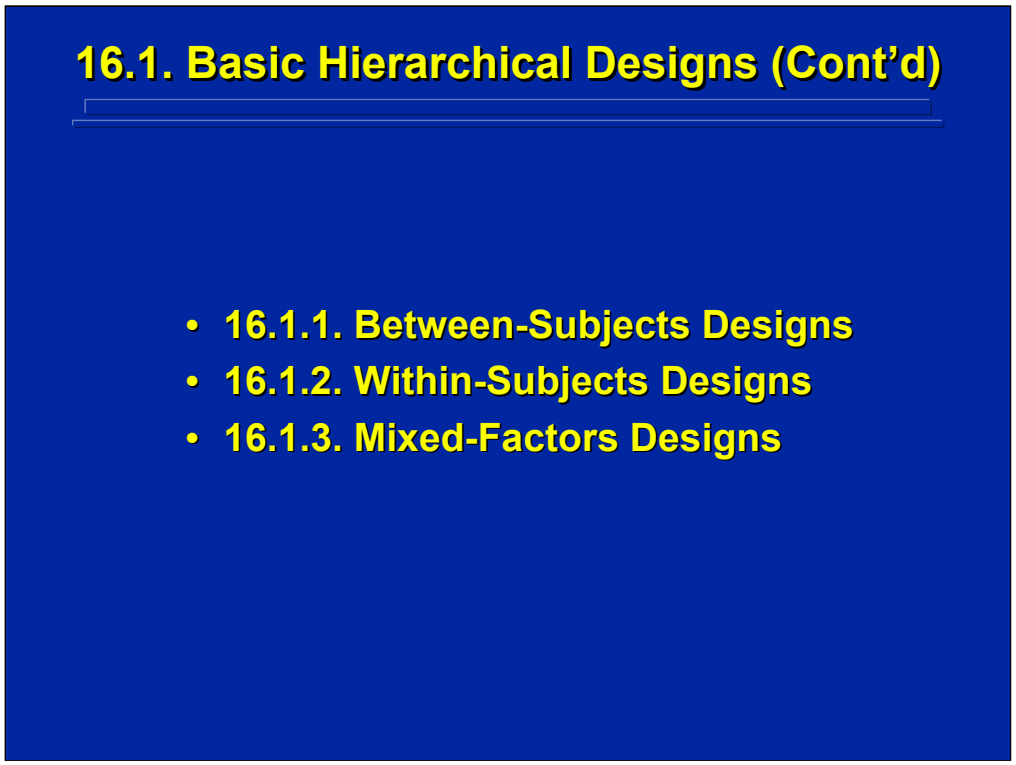



This slide shows a diagram of a three-factor, complete hierarchical design where sample size is 10. A total of 80 observations are in this hierarchical design experiment, which can be conducted as a between-subjects, within-subjects, or mixed-factors design. Note that two levels of Factor B are nested in each level of Factor A, and two levels of Factor C are nested in each of the AB combinations to yield a total of 8 treatment combinations (i.e. cells) in this complete hierarchical design. If this design were a completely crossed factorial design, there would be 64 cells in the 2x4x8 factorial as compared to 8 cells in the three-factor, complete hierarchical design.

Remember when factors are nested, they cannot interact. Therefore, only main effects and no interactions can be evaluated in a complete hierarchical design. In addition, some of the main effects only exist as nested effects in a hierarchical design. In this three-factor, complete hierarchical design, one can estimate Factor A, Factor B nested within A (i.e., B/A) and Factor C nested within both Factors A and B (i.e., C/AB).




This slide shows a three-factor partial hierarchical design that also has 80 observations across the 8 cells of the design with n = 10. Note that two levels of Factor B are nested within each level of Factor A. But, Factor C is crossed with both Factors A and B. The main effects in this design exist as A, B/A, and C. Due to the factor nesting relationship, only the CxA and the CxB/A two-way interactions exist. Again this partial hierarchical design can be conducted as a between-subjects, within-subject, or mixed factors design depending on the crossing and nesting of subjects with the factors of interest in the experiment.




All the rules, procedures, and algorithms described in basic ANOVA completely crossed factorial designs as described in Section 3 apply to hierarchical designs. Depending on the assignment of subjects to treatment conditions the hierarchical design can be conducted as a between-subjects, within-subjects, or mixed factors design. Each of these three design categories are considered separately for both the three-factor complete and partial hierarchical design examples.

## 16.1.1. Between-Subjects Designs

### Between-Subjects, Complete Hierarchical Design

$$Y_{ijklm} = \mu + \alpha_i + \beta_{j(i)} + \delta_{k(ij)} + \gamma_{l(ijk)} + \varepsilon_{m(ijkl)}$$

| Source | df | E(MS) |
|---|---|---|
| A | a-1 = 1 | $bcn\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| B/A | a(b-1) = 2 | $cn\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| C/AB | ab(c-1) = 4 | $n\sigma_\delta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| S/ABC | abc(n-1) = 72 | $\sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| Total | abcn-1 = 79 | |

First, consider the complete hierarchical design as a between-subjects design. Ten different subjects would be observed in each cell of the design matrix shown on page 494, or a total of 80 different subjects are needed for this between-subjects experiment. The statistical model is depicted on the top of this slide showing the nesting relationships in parenthesis of Factor B ($\beta$), Factor C ($\delta$), and Subjects ($\gamma$). Because of this nesting relationship, only main effects and no interactions appear in the statistical model and the resulting listing of Sources.

It is important to note that the degrees of freedom of nested factors are determined by the number of levels nested not the total number of different levels of that factor appearing in the experiment. Hence, b = 2 not 4, and c = 2 not 8 just as n = 10 not 80 for subjects. The standard rules for determining degrees of freedom can be used to specify the df of each effect as shown in the slide.

The expected mean squares follow the same algorithm as used in basic ANOVA. Based on the resulting E(MS) listing shown on this slide, S/ABC is the appropriate error term for testing the three main effects in this between-subjects, complete hierarchical design.

## 16.1.1. Between-Subjects Designs (Cont'd)

### Between-Subjects, Partial Hierarchical Design

$$Y_{ijklm} = \mu + \alpha_i + \beta_{j(i)} + \delta_k + \gamma_{l(ijk)} + \alpha\delta_{ik} + \beta\delta_{kj(i)} + \varepsilon_{m(ijkl)}$$

| Source | df | E(MS) |
|---|---|---|
| A | a-1 = 1 | $bcn\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| B/A | a(b-1) = 2 | $cn\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| C | c-1 = 1 | $abn\sigma_\delta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| CxA | (a-1)(c-1) = 1 | $bn\sigma_{\alpha\delta}^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| CxB/A | a(b-1)(c-1) = 2 | $n\sigma_{\beta\delta}^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| S/ABC | abc(n-1) = 72 | $\sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| Total | abcn-1 = 79 | |

This slide depicts the statistical model, Sources, df, and E(MS) of the between-subjects version of the partial hierarchical design in which only B is nested within A (i.e., B/A). Again, ten different subjects would be observed in each cell of the design matrix shown on page 495, or a total of 80 different subjects are needed for this between-subjects experiment. Note that b =2 not 4 when determining the df of any B/A effects in the design. Due to the nesting relationship, both the CxA and the CxB/A two-way interactions can be evaluated in this design as compared to the complete hierarchical design alternative where no interactions can be evaluated. As in all between-subjects designs, the E(MS) listing demonstrates that S/ABC is the appropriate error term to test every effect assuming all factors of interest are fixed-effect factors.

## 16.1.2. Within-Subjects Designs

### Within-Subjects, Complete Hierarchical Design

$$Y_{ijklm} = \mu + \alpha_i + \beta_{j(i)} + \delta_{k(ij)} + \gamma_l + \alpha\gamma_{il} + \beta\gamma_{j(i)l} + \delta\gamma_{k(ij)l} + \varepsilon_{m(ijkl)}$$

| Source | df | E(MS) |
|---|---|---|
| **Between** | | |
| S | n-1 = 9 | $abc\sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| **Within** | | |
| A | a-1 = 1 | $bcn\sigma_\alpha^2 + bc\sigma_{\alpha\gamma}^2 + \sigma_\varepsilon^2$ |
| AxS | (a-1)(n-1) = 9 | $bc\sigma_{\alpha\gamma}^2 + \sigma_\varepsilon^2$ |
| B/A | a(b-1) = 2 | $cn\sigma_\beta^2 + c\sigma_{\beta\gamma}^2 + \sigma_\varepsilon^2$ |
| B/AxS | a(b-1)(n-1) = 18 | $c\sigma_{\beta\gamma}^2 + \sigma_\varepsilon^2$ |
| C/AB | ab(c-1) = 4 | $n\sigma_\delta^2 + \sigma_{\delta\gamma}^2 + \sigma_\varepsilon^2$ |
| C/ABxS | ab(c-1)(n-1) = 36 | $\sigma_{\delta\gamma}^2 + \sigma_\varepsilon^2$ |
| **Total** | abcn-1 = 79 | |

The statistical model, Source, df, and E(MS) listings for the within-subjects version of the complete hierarchical design are shown on this slide. Note that the same ten subjects appear in all eight cells of the design shown on page 494 and are crossed with all the three factors of interest in the experiment. According to the E(MS) listing, the A, B/A, and C main effects are tested by their interaction with subjects. However, the subject main effect (S) is not tested as in a basic ANOVA within-subjects design.

## 16.1.2. Within-Subjects Designs (Cont'd)

### Within-Subject, Partial Hierarchical Design

$$Y_{ijklm} = \mu + \alpha_i + \beta_{j(i)} + \delta_k + \gamma_l + \alpha\delta_{ik} + \alpha\gamma_{il} + \delta\gamma_{kl} + \beta\delta_{j(i)k} + \beta\gamma_{j(i)l} + \alpha\delta\gamma_{ikl} + \beta\delta\gamma_{j(i)kl} + \varepsilon_{m(ijkl)}$$

| Source | df | E(MS) |
|---|---|---|
| **Between** | | |
| S | n-1 = 9 | $abc\sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| **Within** | | |
| A | a-1 = 1 | $bcn\sigma_\alpha^2 + bc\sigma_{\alpha\gamma}^2 + \sigma_\varepsilon^2$ |
| AxS | (a-1)(n-1) = 9 | $bc\sigma_{\alpha\gamma}^2 + \sigma_\varepsilon^2$ |
| B/A | a(b-1) = 2 | $cn\sigma_\beta^2 + c\sigma_{\beta\gamma}^2 + \sigma_\varepsilon^2$ |
| B/AxS | a(b-1)(n-1) = 18 | $c\sigma_{\beta\gamma}^2 + \sigma_\varepsilon^2$ |
| C | c-1 = 1 | $abn\sigma_\delta^2 + ab\sigma_{\delta\gamma}^2 + \sigma_\varepsilon^2$ |
| CxS | (c-1)(n-1) = 9 | $ab\sigma_{\delta\gamma}^2 + \sigma_\varepsilon^2$ |
| AxC | (a-1)(c-1) = 1 | $bn\sigma_{\alpha\delta}^2 + b\sigma_{\alpha\delta\gamma}^2 + \sigma_\varepsilon^2$ |
| AxCxS | (a-1)(c-1)(n-1) = 9 | $b\sigma_{\alpha\delta\gamma}^2 + \sigma_\varepsilon^2$ |
| B/AxC | a(b-1)(c-1) = 2 | $n\sigma_{\beta\delta}^2 + \sigma_{\beta\delta\gamma}^2 + \sigma_\varepsilon^2$ |
| B/AxCxS | a(b-1)(c-1)(n-1) = 18 | $\sigma_{\beta\delta\gamma}^2 + \sigma_\varepsilon^2$ |
| **Total** | abcn-1 = 79 | |

The statistical model, Source, df, and E(MS) listings for the within-subjects version of the partial hierarchical design are shown on this slide. Again, the same ten subjects appear in all eight cells of this design on page 495 and are crossed with all the three factors of interest in the experiment. According to the E(MS) listing, the A, B/A, and C main effects and the AxC and B/AxC two-way interactions are tested by their interaction with subjects while the subject main effect (S) is not tested as in a basic ANOVA within-subjects design.

## 16.1.3. Mixed-Factors Designs

### Mixed-Factor, Partial Hierarchical Design

$$Y_{ijklm} = \mu + \alpha_i + \beta_{j(i)} + \delta_k + \gamma_{l(ij)} + \alpha\delta_{ik} + \beta\delta_{j(i)k} + \delta\gamma_{kl(ij)} + \varepsilon_{m(ijkl)}$$

| Source | df | E(MS) |
|---|---|---|
| **Between** | | |
| A | a-1 = 1 | $bcn\sigma_\alpha^2 + c\sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| B/A | a(b-1) = 2 | $cn\sigma_\beta^2 + c\sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| S/AB | ab(n-1) = 36 | $c\sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| **Within** | | |
| C | c-1 = 1 | $abn\sigma_\delta^2 + \sigma_{\delta\gamma}^2 + \sigma_\varepsilon^2$ |
| AxC | (a-1)(c-1) = 1 | $bn\sigma_{\alpha\delta}^2 + \sigma_{\delta\gamma}^2 + \sigma_\varepsilon^2$ |
| B/AxC | a(b-1)(c-1) = 2 | $n\sigma_{\beta\delta}^2 + \sigma_{\delta\gamma}^2 + \sigma_\varepsilon^2$ |
| CxS/AB | ab(c-1)(n-1) = 36 | $\sigma_{\delta\gamma}^2 + \sigma_\varepsilon^2$ |
| **Total** | abcn-1 = 79 | |

Mixed-factors hierarchical designs can only exist as partial hierarchical designs because, by definition, some factors are crossed and some factors are nested with subjects. In the mixed-factors, partial hierarchical design used in this slide, Factors A and B/A are between-subjects factors, and Factor C is a within-subjects factor.

The statistical model, Source, df, and E(MS) listings for this mixed-factors version of a three-factor partial hierarchical design are shown on this slide. Since there are ten subjects per cell, a total of 40 different subjects are needed for the four between-subjects treatment combinations in this experimental design as shown on page 495. According to the E(MS) listing, the A and B/A between-subjects effects are tested by S/AB and the C, AxC, and CxB/A effects are tested by the CxS/AB interaction as in a basic ANOVA mixed-factors design.

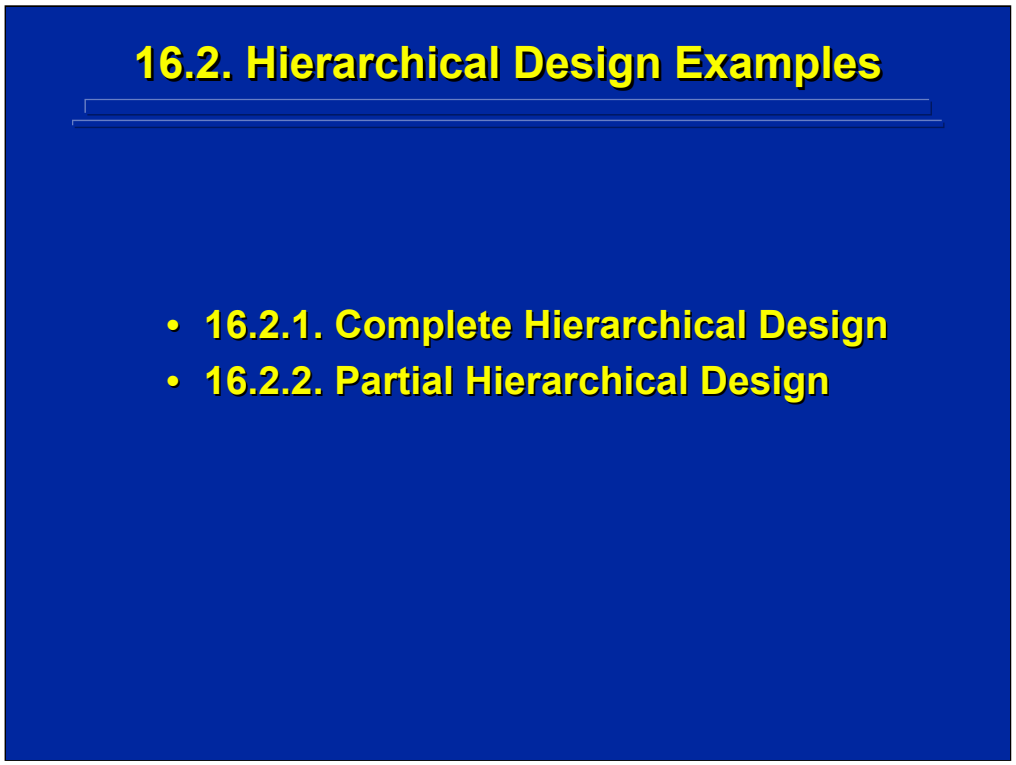
## 16.1.3. Mixed-Factors Design (Cont'd)

### Mixed-Factor, Partial Hierarchical Design

$$Y_{ijklm} = \mu + \alpha_i + \beta_{j(i)} + \delta_k + \gamma_{l(k)} + \alpha\delta_{ik} + \alpha\gamma_{il(k)} + \beta\delta_{j(i)k} + \beta\gamma_{j(i)l(k)} + \varepsilon_{m(ijkl)}$$

| Source | df | E(MS) |
|---|---|---|
| **Between** | | |
| C | c-1 = 1 | $abn\sigma_\delta^2 + ab\sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| S/C | c(n-1) = 18 | $ab\sigma_\gamma^2 + \sigma_\varepsilon^2$ |
| **Within** | | |
| A | a-1 = 1 | $bcn\sigma_\alpha^2 + b\sigma_{\alpha\gamma}^2 + \sigma_\varepsilon^2$ |
| AxC | (a-1)(c-1) = 1 | $bn\sigma_{\alpha\delta}^2 + b\sigma_{\alpha\gamma}^2 + \sigma_\varepsilon^2$ |
| AxS/C | c(a-1)(n-1) = 18 | $b\sigma_{\alpha\gamma}^2 + \sigma_\varepsilon^2$ |
| B/A | a(b-1) = 2 | $cn\sigma_\beta^2 + \sigma_{\beta\gamma}^2 + \sigma_\varepsilon^2$ |
| B/AxC | a(b-1)(c-1) = 2 | $n\sigma_{\beta\delta}^2 + \sigma_{\beta\gamma}^2 + \sigma_\varepsilon^2$ |
| B/AxS/C | ac(b-1)(n-1) = 36 | $\sigma_{\beta\gamma}^2 + \sigma_\varepsilon^2$ |
| Total | abcn-1 = 79 | |

In the mixed-factors, partial hierarchical design used in this slide, Factors A and B/A are within-subjects factors, and Factor C is a between-subjects factor. The statistical model, Source, df, and E(MS) listings for this mixed-factors version of a three-factor partial hierarchical design are shown on this slide. Ten different subjects are needed for each of the two levels of Factor C, and each of these 20 subjects receive all four within-subjects treatment combinations in this experimental design as shown on page 495. According to the E(MS) listing, C is tested by S/C, A and AxC are tested by AxS/C, and B/A and B/AxC are tested by B/AxS/C as in a basic ANOVA mixed-factors design.




Two example between-subjects, hierarchical problems are described in this subsection to demonstrate ANOVA calculations. First, a complete hierarchical design is discussed which is followed by a partial hierarchical design. Both examples use the three-factor design layouts of 80 observations as presented in the previous subsection to facilitate comparisons.

## 16.2.1. Complete Hierarchical Design

- **Example Problem: The military is testing a computer-based multimedia training procedure for commanders. The training procedure is presented to 80 commanders from eight battalions. Two battalions were chosen from each of two brigades within two divisions (infantry and cavalry). The hours to complete the multimedia training on the use of computer-generated surveillance displays were recorded for 10 commanders per battalion. Is training completion time significantly different based on the three command levels? ($p < 0.05$)**

(Click in this red rectangle to see SAS calculations for this example.)

This slide describes a between-subjects, complete hierarchical experimental design problem that has a sample size of 10 (i.e., n = 10). This example problem describes a complete hierarchical design, because battalions (Factor C) are nested within brigades (Factor B), and brigades are nested within divisions (Factor A). Consequently, the 10 battalion commanders represented in each of the eight battalions belong to only one brigade and one particular division resulting in only eight different cells in the experiment. The Slater and Williges (2006) appendix describes the SAS analysis for this example of a complete hierarchical experimental design problem.

## 16.2.1. Complete Hierarchical Design (Cont'd)

### Between-Subjects, Complete Hierarchical Design Example

| Infantry Division | | | | Cavalry Division | | | |
|---|---|---|---|---|---|---|---|
| Brigade 1 | | Brigade 2 | | Brigade 3 | | Brigade 4 | |
| Battalion 1 | Battalion 2 | Battalion 3 | Battalion 4 | Battalion 5 | Battalion 6 | Battalion 7 | Battalion 8 |
| 17 | 29 | 34 | 39 | 23 | 13 | 15 | 15 |
| 28 | 35 | 23 | 21 | 17 | 24 | 25 | 27 |
| 16 | 33 | 34 | 10 | 36 | 11 | 30 | 25 |
| 13 | 29 | 33 | 18 | 21 | 19 | 32 | 18 |
| 31 | 37 | 18 | 23 | 12 | 20 | 40 | 20 |
| 27 | 25 | 26 | 17 | 28 | 33 | 28 | 28 |
| 23 | 32 | 12 | 34 | 32 | 22 | 33 | 11 |
| 16 | 13 | 27 | 39 | 24 | 14 | 16 | 22 |
| 34 | 26 | 24 | 33 | 17 | 19 | 39 | 13 |
| 12 | 29 | 19 | 35 | 36 | 36 | 32 | 25 |

(Click in this red rectangle to see SAS calculations for this example.)

Hypothetical example data representing a battalion commander's time, in hours, to complete multimedia training using computer-generated surveillance displays are presented in the data matrix shown on this slide. Note that the data matrix shows the complete hierarchical relationship of battalions nested in brigades and divisions as well as brigades nested in divisions such that two battalion commanders are nested within each brigade and two brigades are nested within each division.

## 16.2.1. Complete Hierarchical Design (Cont'd)

**Between-Subjects, Complete Hierarchical Design Example**

**Sum of Squares Formulae**

$$SS_A = (\Sigma A^2_{i...}/bcn) - (T^2_{....}/abcn)$$

$$SS_{B/A} = (\Sigma AB^2_{ij..}/cn) - (\Sigma A^2_{i...})/bcn)$$

$$SS_{C/AB} = (\Sigma ABC^2_{ijk.}/n) - (\Sigma AB^2_{i...}/cn)$$

$$SS_{S/ABC} = \Sigma ABCS^2_{ijkn} - (\Sigma ABC^2_{ijk,})/n)$$

$$SS_{Total} = \Sigma ABCS^2_{ijkn} - (T^2_{....})/abcn)$$

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the SS computational formulae for the between-subjects, complete hierarchical design example. The algorithm for determining SS computational formulae described in Topic 10 for basic ANOVA can be used to generate these formulae in simplified notation where Factor A represents divisions; Factor B represents brigades; and Factor C represents battalions.

## 16.2.1. Complete Hierarchical Design (Cont'd)

**Between-Subjects, Complete Hierarchical Design Example**

**Component Scores**

$(\Sigma A^2_{i...}/bcn) = (1024^2 + 921^2)/(2)(2)10) = 488824.4.25$

$(\Sigma AB^2_{ij..}/cn) = (505^2 + 519^2 + \ldots + 494^2)/(2)(10) = 48863.55$

$(\Sigma ABC^2_{ijk.}/n) = (217^2 + 288^2 + \ldots + 204^2)/10 = 49564.7$

$\Sigma ABCS^2_{ijkn} = 17^2 + 28^2 + 16^2 + 13^2 + \ldots + 25^2 = 54163.0$

$(T^2_{....})/abcn) = 1795^2/(2)(2)(2)(10) = 48757.8125$

**Sum of Squares Calculations**

$SS_A = 48824.425 - 48757.8125 = 66.61$

$SS_{B/A} = 48863.55 - 48824.425 = 39.125$

$SS_{C/AB} = 49564.7 - 48863.55 = 701.15$

$SS_{S/ABC} = 54163.0 - 49564.7 = 4598.3$

$SS_{Total} = 54163.0 - 48757.8125 = 5405.1875$

(Click in this red rectangle to see SAS calculations for this example.)

The top portion of this slide uses the hypothetical data to calculate the values of each of the five components that make up the SS formulae provided on the previous slide. Note that a, b, and c each equal 2 to represent the number of nested levels in this complete hierarchical design.

The SS values for this example are determined by combining the various component scores algebraically according to the SS formulae given on the previous slide. The final calculations for A, B/A, C/AB, S/ABC, and Total SS are shown on the bottom portion of this slide.

## 16.2.1. Complete Hierarchical Design (Cont'd)

- ANOVA Summary Table

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Division (D) | 1 | 66.61 | 66.61 | 1.04 |
| Brigade (Br)/D | 2 | 39.13 | 19.57 | 0.31 |
| Battalion (Ba)/DBr | 4 | 701.15 | 175.29 | 2.74* |
| S/DBrBa | 72 | 4598.30 | 63.87 | |
| Total | 79 | 5405.19 | | |

*p < 0.05

(Click in this red rectangle to see SAS calculations for this example.)

The ANOVA Summary Table for the complete hierarchical design example problem is shown on this slide. The Summary Table uses real-world abbreviations such that D = A, Br = B, and Ba = C in the computation formulae using simplified notation.

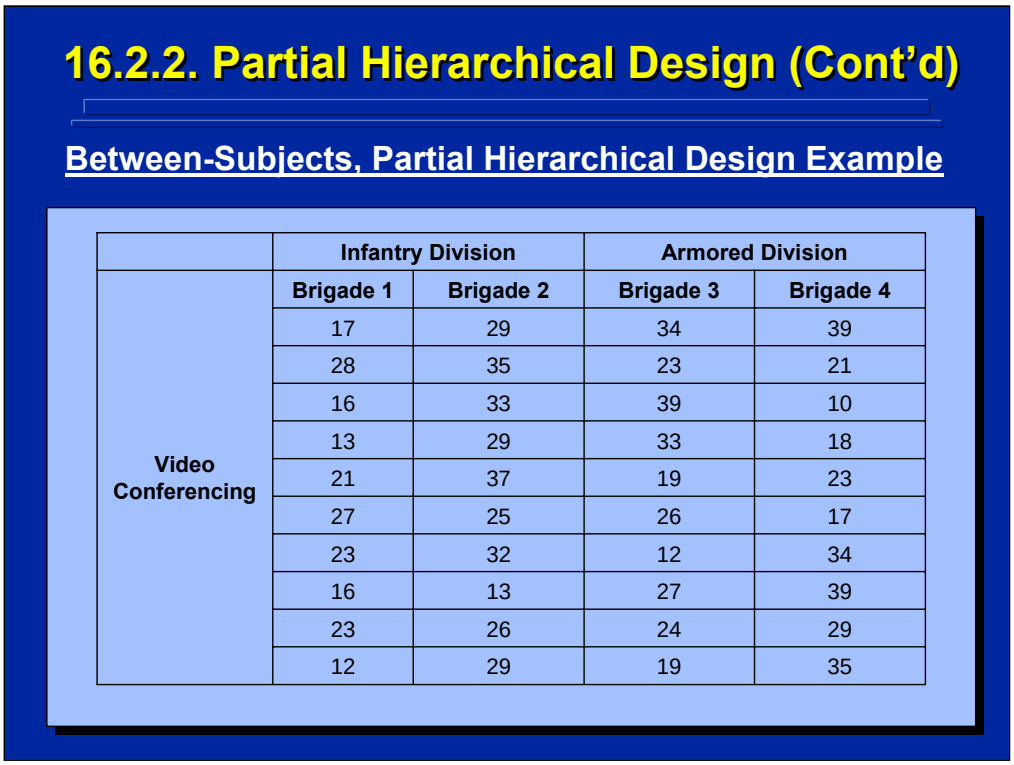
There is a significant difference ($p < 0.05$) in time to complete multimedia training on using computer-based surveillance displays among the battalion command level nested within brigades and divisions. Post hoc tests are needed to isolate the differences among the battalions. Due to the nesting relationships of the command levels, this significant effect could be due either to specific battalion command structures or the interaction of battalion, brigade, and division command levels.

## 16.2.2. Partial Hierarchical Design

- **Example Problem: The military is testing two communication systems used by commanders of four brigades. Two brigades came from an infantry division and two from an armored division. A video conferencing or an instant messaging system was presented to 10 commanders in each brigade. Each commander used only one of the communication systems. The commanders' satisfaction ratings for the systems were recorded. Is there a significant satisfaction difference ($p < 0.05$) between the two communication systems and/or the nesting of commander levels?**

(Click in this red rectangle to see SAS calculations for this example.)

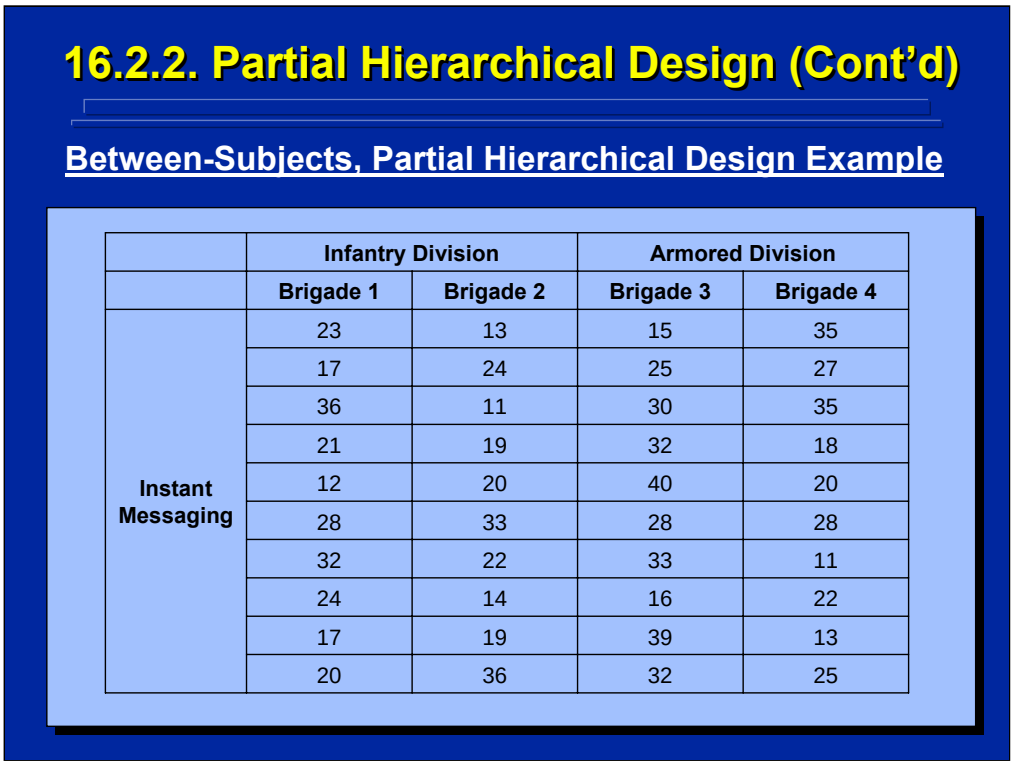
This slide describes a between-subjects, partial hierarchical design involving three factors. Two of the four brigade commanders tested are nested within the two divisions. Command levels are crossed with the two communication systems. Since each brigade commander used either video conferencing or instant messaging, this is a between-subjects, partial hierarchical design with eight treatment combinations and sample size, n, equals 10.




| | Infantry Division | | Armored Division | |
|---|---|---|---|---|
| | Brigade 1 | Brigade 2 | Brigade 3 | Brigade 4 |
| Video Conferencing | 17 | 29 | 34 | 39 |
| | 28 | 35 | 23 | 21 |
| | 16 | 33 | 39 | 10 |
| | 13 | 29 | 33 | 18 |
| | 21 | 37 | 19 | 23 |
| | 27 | 25 | 26 | 17 |
| | 23 | 32 | 12 | 34 |
| | 16 | 13 | 27 | 39 |
| | 23 | 26 | 24 | 29 |
| | 12 | 29 | 19 | 35 |

(Click in this red rectangle to see SAS calculations for this example.)

Hypothetical example of satisfaction ratings of video conferencing communications for the ten commanders from the two brigades nested in the infantry division and the ten commanders from the two brigades nested within the armored division are listed on this slide. This represents the first half of the partial hierarchical design data matrix.




(Click in this red rectangle to see SAS calculations for this example.)

Hypothetical example data of satisfaction ratings of instant messaging communications for the ten commanders from the two brigades nested in the infantry division and the ten commanders from the two brigades nested within the armored division are listed on this slide. This represents the second half of the partial hierarchical design data matrix.

## 16.2.2. Partial Hierarchical Design (Cont'd)

**Between-Subjects, Partial Hierarchical Design Example**

**Sum of Squares Formulae**

$SS_A = (\Sigma A^2_{i...}/bcn) - (T^2_{....}/abcn)$

$SS_{B/A} = (\Sigma AB^2_{ij..}/cn) - (\Sigma A^2_{i...})/bcn)$

$SS_C = (\Sigma C^2_{..k.}/abn) - (T^2_{....}/abcn)$

$SS_{CxA} = (\Sigma AC^2_{i.k.}/bn) - (\Sigma A^2_{i...})/bcn) - (\Sigma C^2_{..k.}/abn) + (T^2_{....}/abcn)$

$SS_{CxB/A} = (\Sigma ABC^2_{ijk.}/n) - (\Sigma AB^2_{ij..})/cn) - (\Sigma AC^2_{i.k.}/bn) + (A^2_{i...}/bcn)$

$SS_{S/ABC} = \Sigma ABCS^2_{ijkn} - (\Sigma ABC^2_{ijk,})/n)$

$SS_{Total} = \Sigma ABCS^2_{ijkn} - (T^2_{....}/abcn)$

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the SS computational formulae for the between-subjects, complete hierarchical design example. The algorithm for determining SS computational formulae described in Topic 10 of this reference material can be used to generate these formulae in simplified notation where Factor A represents divisions; Factor B represents brigades; and Factor C represents communication system.

## 16.2.2. Partial Hierarchical Design (Cont'd)

**Between-Subjects, Partial Hierarchical Design Example**

**Component Scores**

$(\Sigma A^2_{i...}/bcn) = (925^2 + 1045^2)/(2)(2)(10) = 48691.25$

$(\Sigma C^2_{..k.}/abn) = (1005^2 + 965^2)/(2)(2)(10) = 48531.25$

$(\Sigma AB^2_{ij..}/cn) = (426^2 + 499^2 + \ldots + 499^2)/(2)(10) = 48879.70$

$(\Sigma AC^2_{i.k.}/bn) = (484^2 + 521^2 + \ldots + 524^2)/(2)(10) = 48737.70$

$(\Sigma ABC^2_{ijk.}/n) = (196^2 + 288^2 + \ldots + 234^2)/(10) = 49339.80$

$\Sigma ABCS^2_{ijkn} = (17^2 + 28^2 + \ldots + 25^2) = 53874.00$

$(T^2_{....}/abcn) = (1970^2)/(2)(2)(2)(10) = 48511.25$

(Click in this red rectangle to see SAS calculations for this example.)

This slide uses the hypothetical data to calculate the values of each of the seven components that make up the SS formulae provided on the previous slide. Note that a, b, and c each equal 2 to represent the nesting relationship of command levels and the two communication system alternatives in this example.

## 16.2.2. Partial Hierarchical Design (Cont'd)

**Between-Subjects, Partial Hierarchical Design Example**

**Sum of Squares Calculations**

$SS_A = 48691.25 - 48511.25 = 180.00$

$SS_{B/A} = 48879.70 - 48691.25 = 188.45$

$SS_C = 48531.25 - 48511.25 = 20.00$

$SS_{CxA} = 48737.70 - 48691.25 - 48531.25 + 48511.25 = 26.45$

$SS_{CxB/A} = 49339.80 - 48879.70 - 48737.70 + 48691.25 = 413.65$

$SS_{S/ABC} = 53874.00 - 49339.8 = 4534.20$

$SS_{Total} = 53874.00 - 48511.25 = 5362.75$

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows that the SS values for this example are determined by combining the various component scores algebraically according to the SS formulae given on the previous two slides. The final calculations for A, B/A, C, CxA, CxB/A, S/ABC, and Total SS are shown.

## 16.2.2. Partial Hierarchical Design (Cont'd)

- ANOVA Summary Table

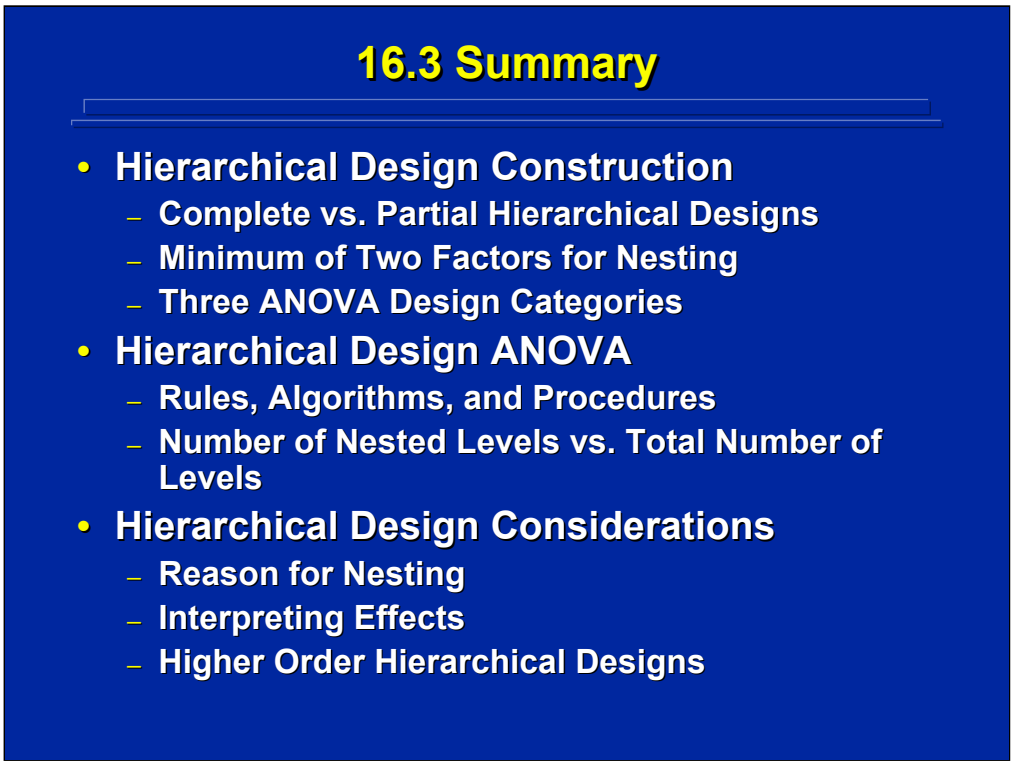
| Source | df | SS | MS | F |
|---|---|---|---|---|
| Division (D) | 1 | 180.00 | 180.00 | 2.86 |
| Brigade (B)/D | 2 | 188.45 | 94.23 | 1.50 |
| Communication (C) | 1 | 20.00 | 20.00 | 0.32 |
| CxD | 1 | 26.45 | 26.45 | 0.42 |
| CxB/D | 2 | 413.65 | 206.83 | 3.28* |
| S/DBC | 72 | 4534.20 | 62.98 | |
| Total | 79 | 5362.75 | | |

*p < 0.05

(Click in this red rectangle to see SAS calculations for this example.)

The ANOVA Summary Table for the partial hierarchical design example problem is shown on this slide. The Summary Table uses real-world abbreviation such that D = A, B = B, and C = C in the computation formulae using simplified notation.

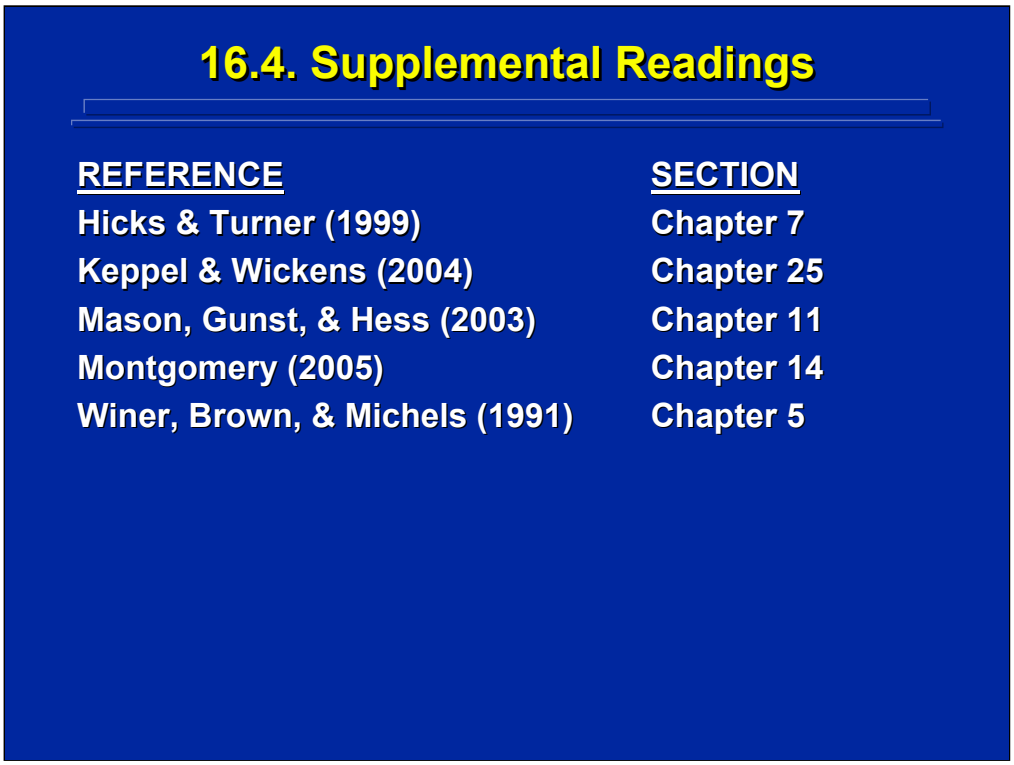
There is a significant difference ($p < 0.05$) in the two-way interaction between communication systems and brigade commanders nested within divisions. Post hoc tests are needed to isolate the difference in this interaction. Due to the nesting relationships of the command levels, the brigade commander component in the interaction could be due to specific brigade commander differences or brigade by division differences or both since brigade level command is nested within infantry and armored divisions.




This topic focused on the construction and analysis of hierarchical designs. In general, complete and partial hierarchical ANOVA designs are determined by the nesting relationships of the factors of interest. Remember that the experimenter must always use a minimum of two levels for nesting. If only one level of a factor is nested within another factor, the factors become confounded not nested. Complete hierarchical designs can be conducted as either between-subjects or within-subjects designs. Partial hierarchical designs can be conducted as between-subjects, within-subjects, or mixed-factors designs.

All the rules, algorithms, and procedures from basic ANOVA apply to hierarchical designs. When calculating the various df and SS in the ANOVA, the number of levels of a nested factor always equal the number of levels nested not the total number of levels of that factor.

In summary, hierarchical designs are used primarily in human factors and ergonomics research when the factors of interest exist as nested factors in the real world and cannot be crossed. Due to the nesting relationship, interpretation of significant effects is problematic due to the confounding of main effects and interactions of nested factors. Higher-order hierarchical designs can easily be constructed but are usually not considered due to the difficulties of interpretation.




All of the chapters in these texts provide a discussion of hierarchical designs used in ANOVA.

# Topic 17. Blocking ANOVA Designs

17.1. Modular Representation
- 17.1.1. Modular Arithmetic
- 17.1.2. Balanced Sets of Treatments
- 17.1.3. Component SS Formulae
- 17.1.4. Generalizations

17.2. Blocking $2^k$ Designs
- 17.2.1. Simple Blocking of $2^k$ Design
- 17.2.2. Complex Blocking of $2^k$ Design
- 17.2.3. Computational Considerations

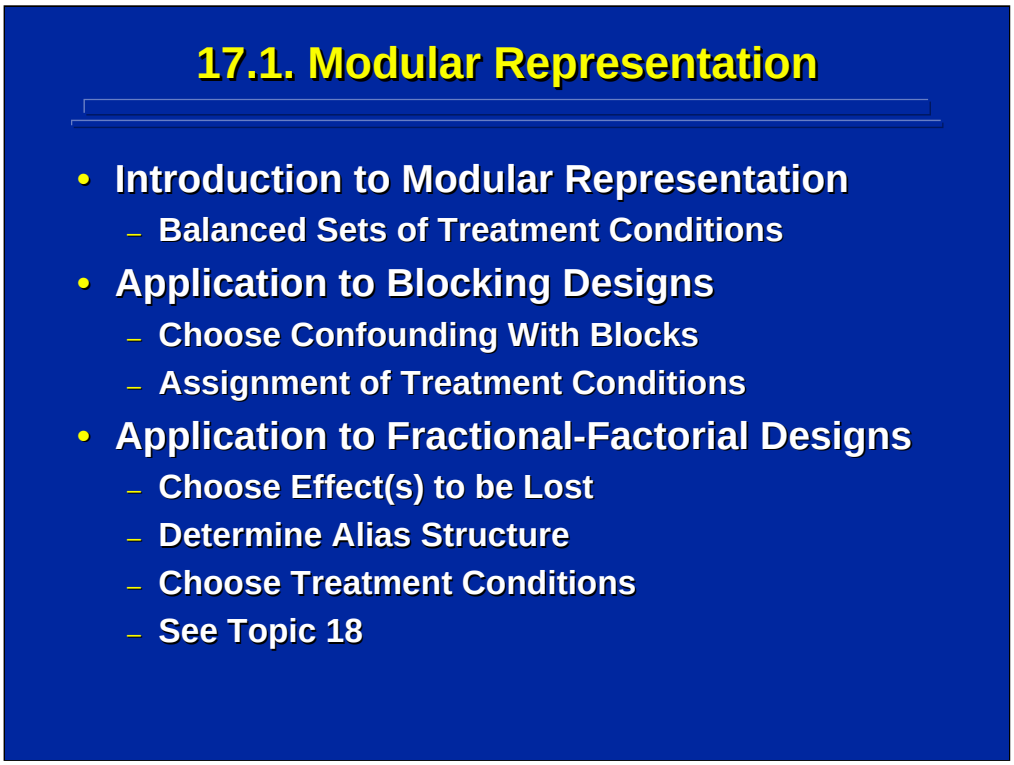
17.3. Pseudo-Factor Blocking

17.4. Summary

17.5. Supplemental Readings

This topic deals with ANOVA experimental designs that can be used to control nuisance variables that are confounded with the factors of interest in an experiment. These confounding factors may include the effect of repeated testing sessions or experimenter bias resulting from using several experimenters in data collection. Confounding is controlled through blocking designs where blocks represent the nuisance variable.

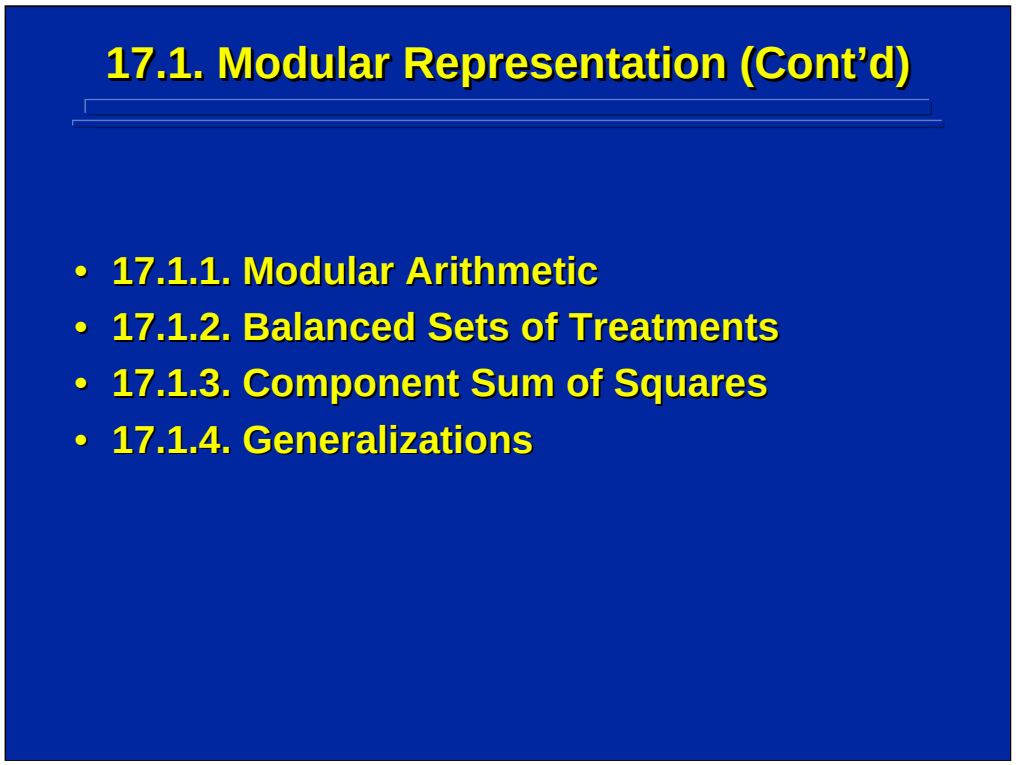
The goal of blocking is to avoid confounding the blocking effect with effects in a factorial ANOVA design that are of major interest to the experimenter. Through the use of modular representation, the exact nature of the block confounding with specific treatment effects of interest can be determined. The use of blocking is demonstrated with $2^k$ factorial designs and extended to pseudo-factor blocking of factors with levels that are not prime numbers. This topic ends with a summary and a list of suggested supplemental readings on blocking in standard experimental design textbooks.



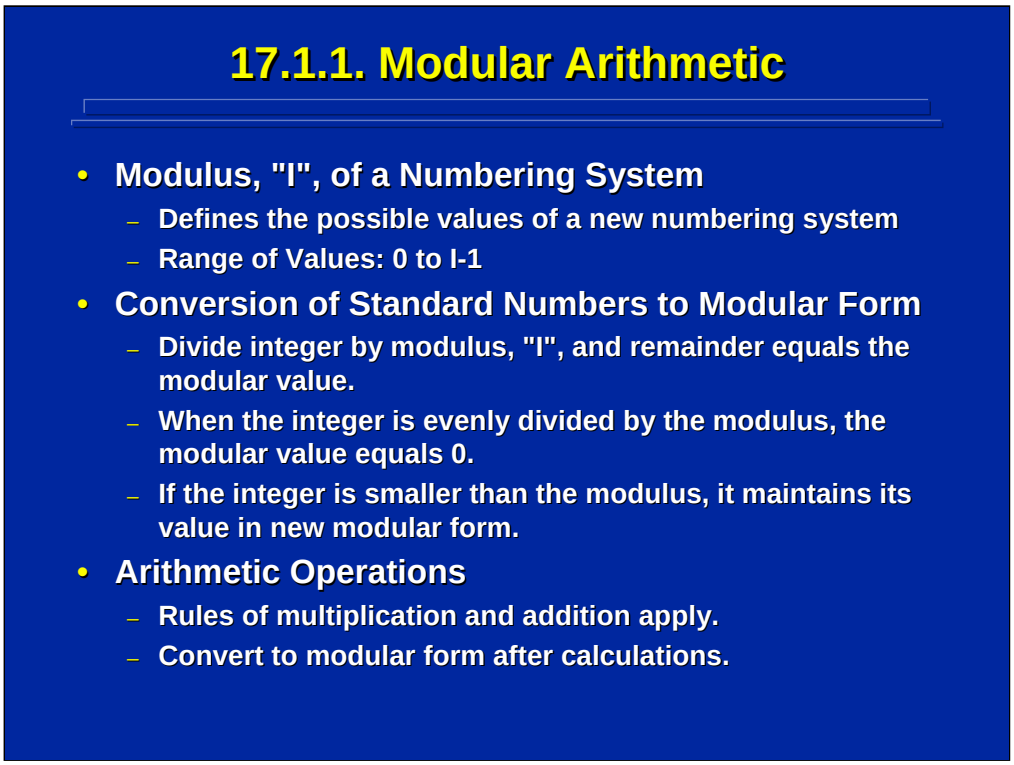

Modular representation, which uses an alternative numbering system, provides the fabric for constructing balanced sets of treatment combinations across various effects in a factorial design. Modular representation can be used to determine the appropriate subset of data to collect to avoid confounding effects of primary interest to the experimenter. These balanced sets can be used to construct both blocking ANOVA designs and fractional-factorial ANOVA designs.

This topic focuses on the use of modular representation to construct blocking designs. For example, a 2x2x2 factorial design may need to be divided into two days of data collection due to the size of the design. The experimenter can use modular representation to choose the four treatment conditions collected each day such that only the three-way interaction of the factorial design is confounded with data collection days (i.e., blocks).

Modular representation can also be used to choose only a fractional subset of a large experimental design for data collection when it is impractical to use a complete factorial design. This is called fractional factorial design or factional replicate of the full factorial design. Fractional replicates are not able to test some effects and confound other effects (i.e., the alias structure) in the full factorial design. Fractional factorials are discussed in Topic 18.

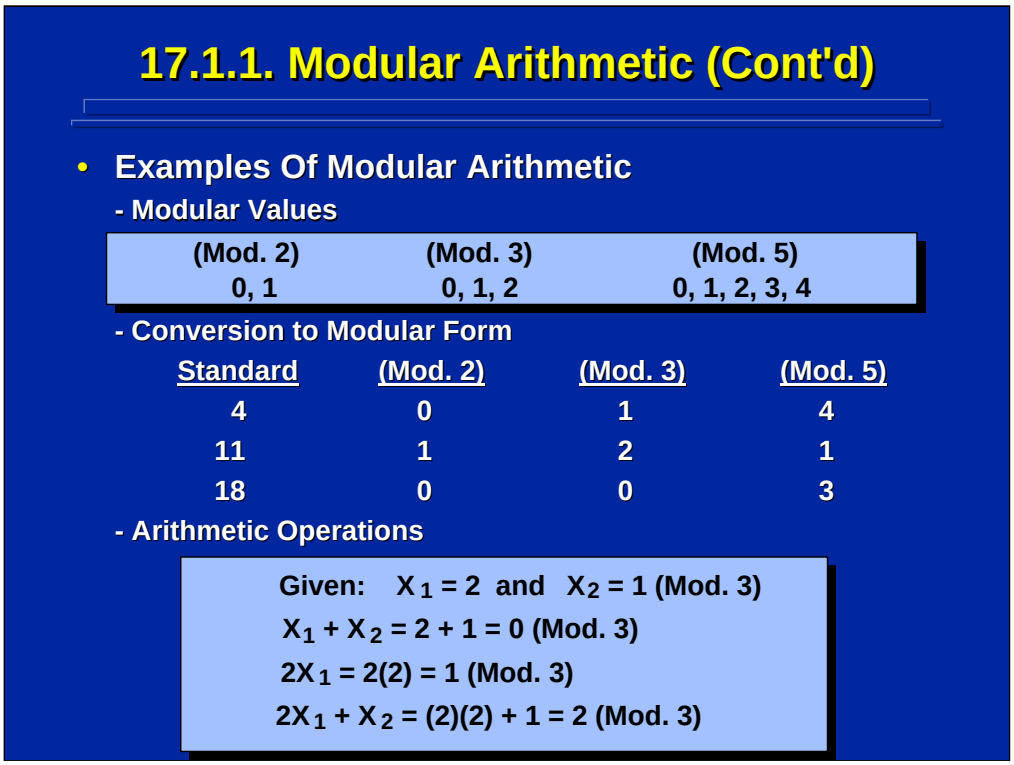



The basic rules of modular arithmetic needed in applications of modular representation to experimental design are reviewed first. Next, modular representation is used to generate balanced sets of treatment conditions, and formulae are presented for calculating the component sum of squares for these balanced sets. Finally, modular representation is generalized to $2^k$, $3^k$, and $5^k$ factorial designs.




Modular arithmetic is used to convert a number in the standard base 10 system to a new numbering system that has a different base. Each numbering system can be defined in terms of a base value or modulus, I, where values range from 0 to I-1. To convert the standard number to the new numbering system, simply divide the old standard value by the new modulus and the remainder will be the new modulus value. If there is no remainder, then the new value is zero.

Arithmetic operations can be conducted in the new modular numbering system just as in the standard numbering system. All rules of arithmetic are applied to the numbers first and then converted back to the original modulus system.



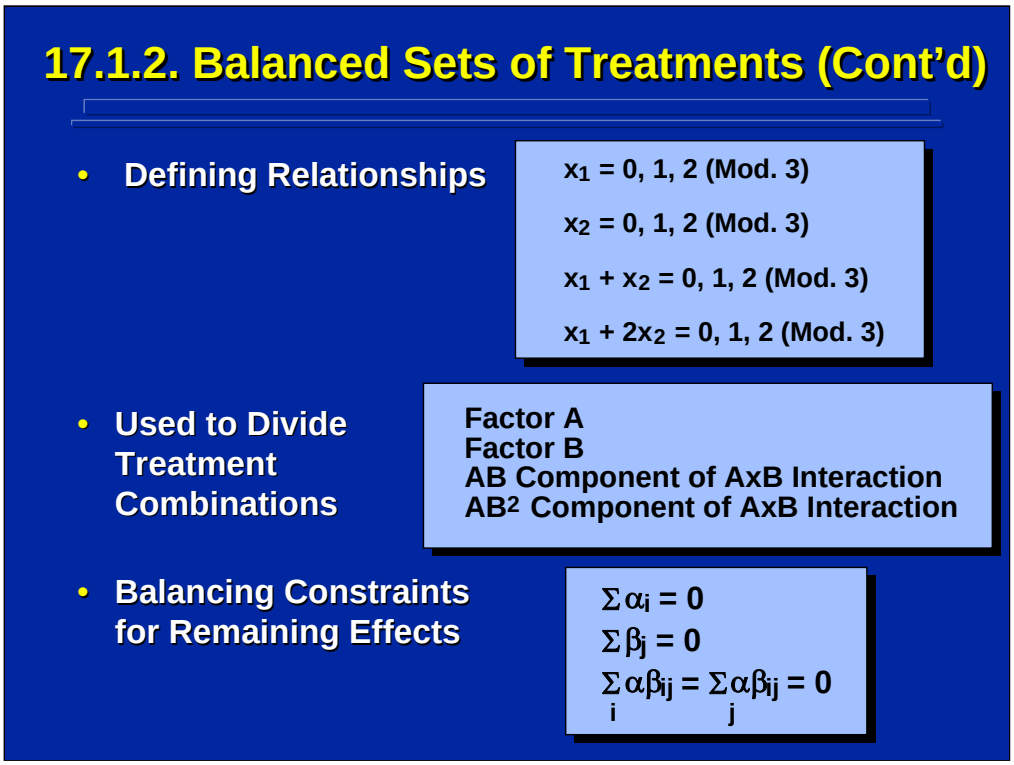

This slide shows examples of using modular arithmetic. Any number in the 10-based numbering system can be converted to a new modular system by dividing by the modulus and using 0 for equal division or the remainder as the new value. Examples in the middle portion of this slide are provided for converting the numbers 4, 11, and 18 to modulus 2 (Mod. 2), modulus 3 (Mod. 3) and modulus 5 (Mod. 5) systems. The bottom portion of this slide shows various addition and multiplication operations in a Mod. 3 system. Any arithmetic operation can be conducted, and the result is then converted to the modulus value.

## 17.1.2. Balanced Sets of Treatments

- **Constraints**
  - Factors are fixed-effects
  - Factor levels are prime numbers
- **Representation of Treatment Conditions**

| Standard Representation | | | | Modular Representation | | | |
|---|---|---|---|---|---|---|---|
| | $a_1$ | $a_2$ | $a_3$ | | $a_0$ | $a_1$ | $a_2$ |
| $b_1$ | $a_1b_1$ | $a_2b_1$ | $a_3b_1$ | $b_0$ | (00) | (10) | (20) |
| $b_2$ | $a_1b_2$ | $a_2b_2$ | $a_3b_2$ | $b_1$ | (01) | (11) | (21) |
| $b_3$ | $a_1b_3$ | $a_2b_3$ | $a_3b_3$ | $b_2$ | (02) | (12) | (22) |

Modular representation is used to construct balanced sets of treatments in blocking and fractional-factorial designs when the levels of the factor all have the same prime number and the factors are fixed-effects factors. The number of levels in these designs determine the modulus used. For example, $2^k$ factorial designs use Mod. 2 representations and $3^k$ factorial designs use Mod. 3 representations. This slide depicts the treatment levels of a 3x3 factorial design in standard representation on the left side of the slide and Mod. 3 representation on the right side of the slide.




Defining relationships in modular notation can be used to divide a factorial design into balanced sets of treatment combinations. These sets represent blocks in blocking designs and each set represents a partial replicate in a fractional-factorial design.

This slides provides a demonstration of using various Mod. 3 defining relationships for the various 2 df components of a 3x3 factorial design. Each 2 df component of this design can be defined by one of the four defining relationships given at the top of this slide where $X_1$ is the level of Factor A, $X_2$ is the level of Factor B, and $X_1 + X_2$ and $X_1 + 2X_2$ define the two 2 df orthogonal components of the AxB interaction in Mod. 3. Three of the nine treatment conditions in the 3x3 factorial design represent each of the values 0,1, 2 (Mod. 3) in each of the four defining relationships representing the 2 df components of the complete factorial design.

The resulting sets of treatment combinations are confounded with the specific value of the defining relationship, but are balanced across the values of remaining effects in the factorial design. Consequently, they define three groups of three balanced sets of treatment conditions that are balanced, rather than confounded, across the remaining 2 df component effects of Factor A, Factor B, and the AxB interaction as defined in the bottom portion of this slide.

## 17.1.2. Balanced Sets of Treatments (Cont'd)

**Confounding Factor A**

**Defining Relationship: $x_1 = 0, 1, 2$ (Mod. 3)**

| $x_1 = 0$ | $x_1 = 1$ | $x_1 = 2$ |
|---|---|---|
| 00 | 10 | 20 |
| 01 | 11 | 21 |
| 02 | 12 | 22 |
| $A_0$ | $A_1$ | $A_2$ |

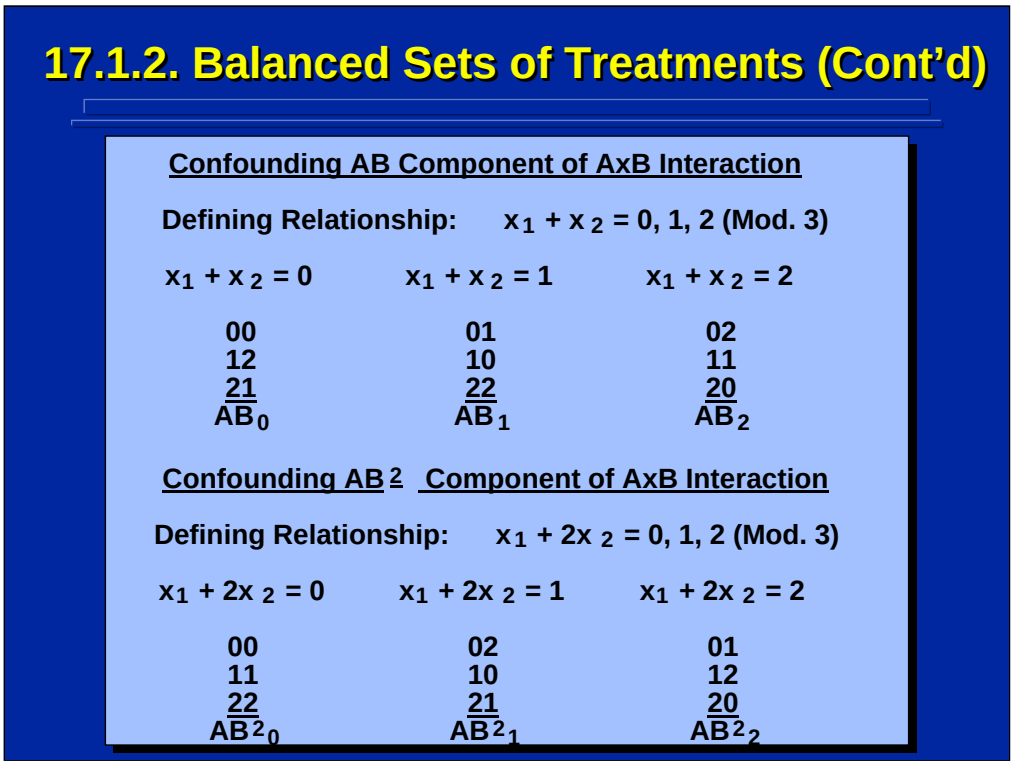
**Confounding Factor B**

**Defining Relationship: $x_2 = 0, 1, 2$ (Mod. 3)**

| $x_2 = 0$ | $x_2 = 1$ | $x_2 = 2$ |
|---|---|---|
| 00 | 01 | 02 |
| 10 | 11 | 12 |
| 20 | 21 | 22 |
| $B_0$ | $B_1$ | $B_2$ |

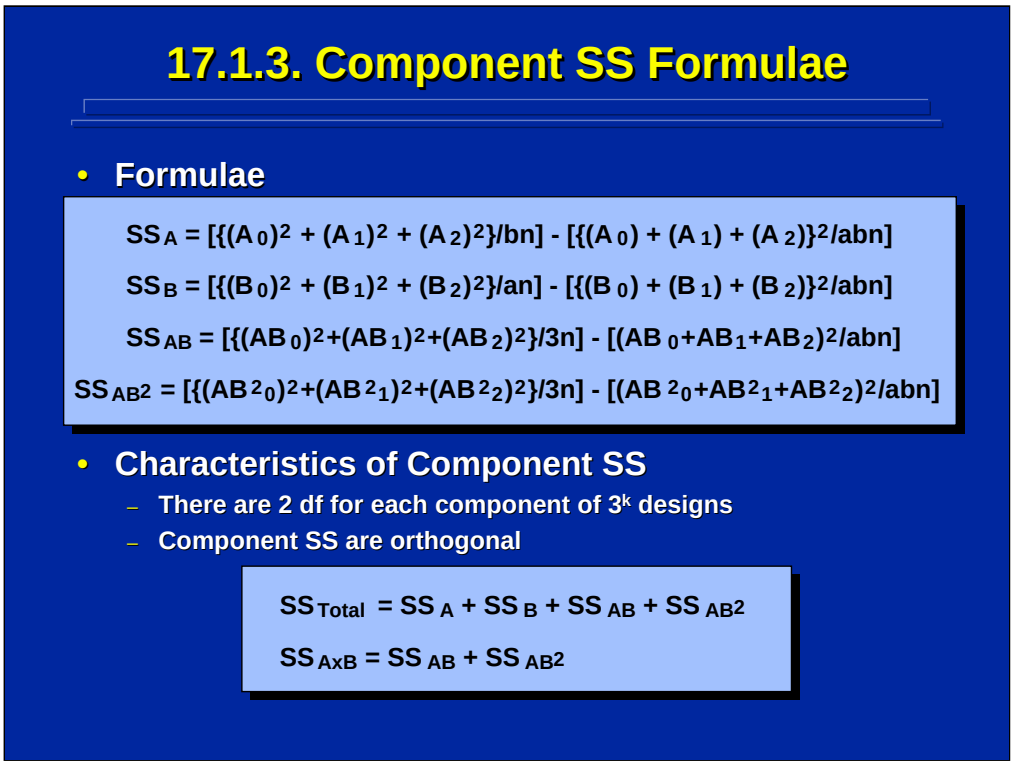
Two different sets of three balanced treatment conditions are shown on this slide. First, the defining relationship shown on the top portion of the slide separates the sets only by levels of Factor A by stating $x_1 = 0, 1, 2$ (Mod. 3). This confounds the three levels of Factor A with the three sets of three treatment conditions shown on the top of this slide, and the remaining effects are balanced across these three sets. Note that Factor A is totally confounded with the three sets of treatment conditions because level 0 of Factor A only appears in the first set, level 1 only appears in the second set, and level 2 only appears in the third set. But, Factor B and the AxB interaction are balanced across the three sets.

Likewise, Factor B is totally confounded with the three sets of treatment conditions shown in the bottom portion of the slide which uses $x_2 = 0, 1, 2$ (Mod. 3) as the defining relationship. But, Factor A and the AxB interaction are balanced across the three sets.




This slide lists the confounding of each of the two orthogonal 2 df components of the AxB interaction with the three sets of balanced treatments. The defining relationship shown in top portion of the slide confounds the AB component of the AxB interaction, and the defining relationship shown in the bottom portion of the slide shows the confounding of the $AB^2$ component of the AxB interaction. The AB and $AB^2$ components are nothing more than two orthogonal components of the AXB interaction. Note that Factors A and B are balanced across both alternatives because all three levels of each factor appear in the three treatments in each of the three balanced sets.

Consequently, the defining relationship determines what is confounded with the balanced sets of treatments. These balanced sets become the blocks in ANOVA blocking designs, and the defining relationship specifies the effect in the factorial design that is confounded with blocks. Obviously, the experimenter would not choose either of the defining relations on the previous slide that confound main effects of either Factor A or B with blocks in this 3x3 factorial design. Rather, confounding one-half of the AxB interaction by choosing either the AB or $AB^2$ component shown on this slide would be a better choice for blocking the example 3x3 factorial design.




The top of this slide shows the formulae that can be used to calculate the four alternative component SS in the 3x3 design described in the previous slides. Each of these components has 2 df. Each of the sum of squares for the components are orthogonal and sum to the Total SS. The SS of the AB and $AB^2$ components sum to the AxB interaction SS because they represent orthogonal 2 df components of the AxB interaction that has 4 df.

## 17.1.3. Component SS Formulae (Cont'd)

- **Two-Factor Design: a = 3, b = 3, and n = 2**
  - **Data Matrix**

| | $a_0$ | $a_1$ | $a_2$ | |
|---|---|---|---|---|
| $b_0$ | 8<br>28 | 15<br>32 | 25<br>42 | $B_0$=150 |
| $b_1$ | 10<br>16 | 16<br>18 | 18<br>10 | $B_1$= 88 |
| $b_2$ | 20<br>22 | 15<br>8 | 25<br>40 | $B_2$=130 |
| | $A_0$=104 | $A_1$=104 | $A_2$=160 | T... = 368 |

This slide shows a hypothetical data matrix (n = 2) of the 3x3 example design. The SS for the AxB interaction can be divided into the two orthogonal components AB and $AB^2$.

## 17.1.3. Component SS Formulae (Cont'd)

- **AxB Interaction Matrix**

| | $a_0$ | $a_1$ | $a_2$ |
|---|---|---|---|
| $b_0$ | (00) = 36 | (10) = 47 | (20) = 67 |
| $b_1$ | (01) = 26 | (11) = 34 | (21) = 28 |
| $b_2$ | (02) = 42 | (12) = 23 | (22) = 65 |

$SS_{AxB} = 358.222$

- **$SS_{AB}$ Component**

| $x_1 + x_2 = 0$ | $x_1 + x_2 = 1$ | $x_1 + x_2 = 2$ |
|---|---|---|
| (00) = 36 | (01) = 26 | (02) = 42 |
| (12) = 23 | (10) = 47 | (11) = 34 |
| (21) = 28 | (22) = 65 | (20) = 67 |
| $AB_0 = 87$ | $AB_1 = 138$ | $AB_2 = 143$ |

$$SS_{AB} = [\{(AB_0)^2+(AB_1)^2+(AB_2)^2\}/3n] - [(AB_0+AB_1+AB_2)^2/abn]$$

$$= [\{(87)^2+(138)^2+(143)^2\}/(3)(2)] - [(87+138+143)^2/(3)(3)(2)]$$

$$SS_{AB} = 320.111$$

The top of this slide lists the $AB_{ij}$ totals used to calculate the AXB interaction in the basic ANOVA computational formula. Using these totals and the totals provided on the previous slide results in the $SS_{AxB}$ = 358.22.

The middle portion of this slide divides the nine treatment combinations into three balanced sets using the AB component of the AxB interaction as the defining relationship, and shows that these three component totals are 87, 138, and 143, respectively.

The bottom portion of this slide shows the calculation of the SS for AB component using the example data. By using the component SS formula, one determines that $SS_{AB}$ = 320.111. Note this value is less than $SS_{AxB}$ (i.e., 358.222) because the AB component represents only 2 df of the total 4 df in the AxB interaction.

## 17.1.3. Component SS Formulae (Cont'd)

- **$SS_{AB^2}$ Component**

| $x_1 + 2x_2 = 0$ | $x_1 + 2x_2 = 1$ | $x_1 + 2x_2 = 2$ |
|---|---|---|
| (00) = 36 | (02) = 42 | (01) = 26 |
| (11) = 34 | (10) = 47 | (12) = 23 |
| (22) = 65 | (21) = 28 | (20) = 67 |
| $AB^2_0$=135 | $AB^2_1$=117 | $AB^2_2$=116 |

$$SS_{AB^2} = [\{(AB^2_0)^2+(AB^2_1)^2+(AB^2_2)^2\}/3n] - [(AB^2_0+AB^2_1+AB^2_2)^2/abn]$$

$$= [\{(135)^2+(117)^2+(116)^2\}/(3)(2)] - [(135+117+116)^2/(3)(3)(2)]$$

$$SS_{AB^2} = 38.111$$

- $SS_{AxB} = SS_{AB} + SS_{AB^2}$

$$358.222 = 320.111 + 38.111$$

The top and middle portions of this slide divide the nine treatment combinations into three balanced sets using the $AB^2$ component of the AxB interaction as the defining relationship, and shows that these three component totals are 135, 117, and 116, respectively. By using the component SS formula, one determines that $SS_{AB}{}^2 = 38.111$. Note this value is less than $SS_{AxB}$ (i.e., 358.222) because the AB component represents only 2 df of the total 4 df in the AxB interaction.

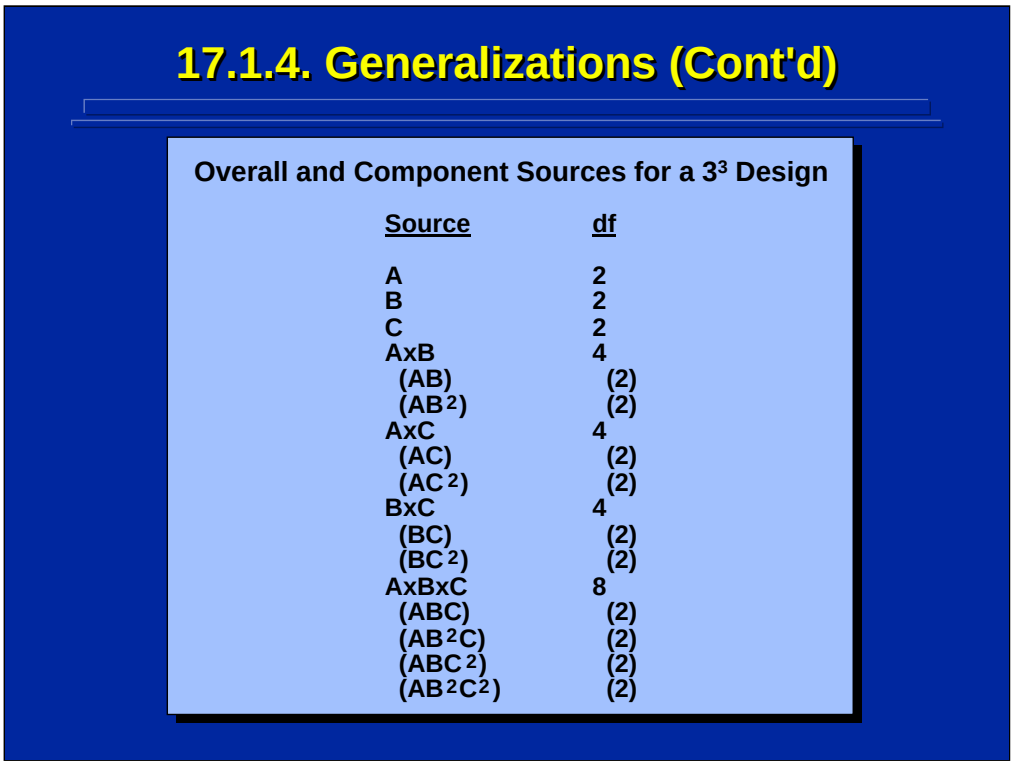
The bottom portion of this slide demonstrates that the two orthogonal components of the two-way interaction (i.e., AB and $AB^2$) sum to the SS originally calculated for the AxB interaction (i.e., 358.222) since the two components are orthogonal. The experimenter has no way of knowing beforehand which of these two components will be larger only that they are orthogonal 2 df component sets of treatment conditions.

## 17.1.4. Generalizations

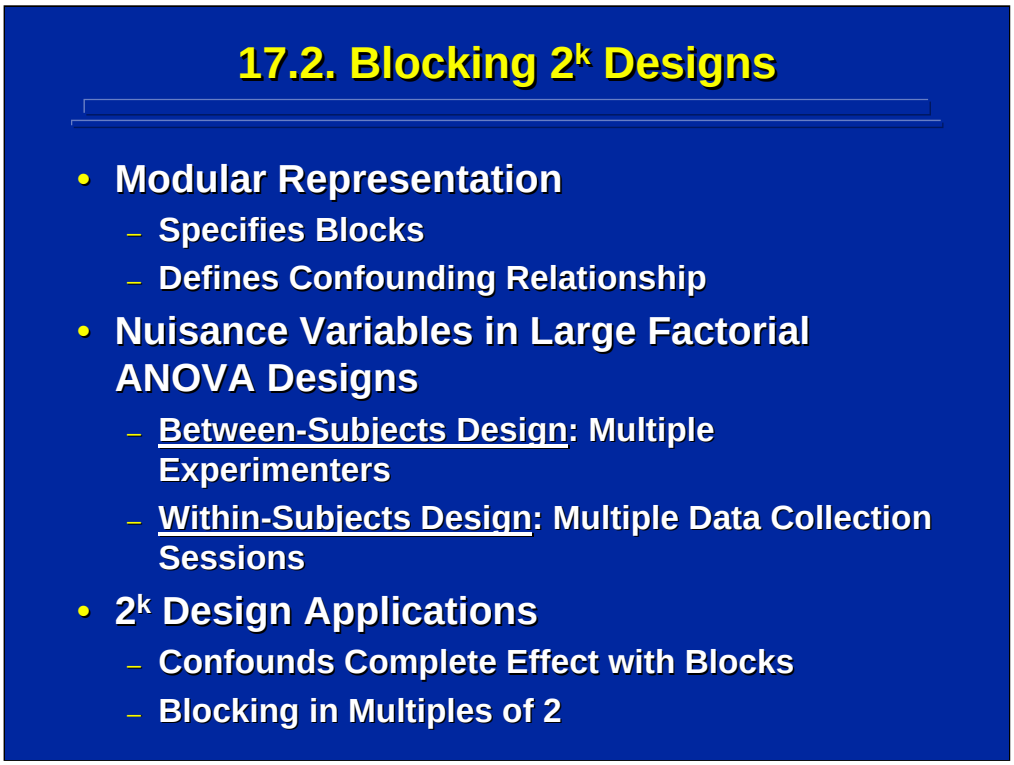
- **Use Primarily $2^k$, $3^k$, and $5^k$ Designs**
- **Orthogonal Interaction Components of $3^k$ and $5^k$ Designs**
  - **Rule: Orthogonal components for the various interactions are determined by maintaining Factor A at a weighting of one and exhausting all the non-zero modular weightings of the other factor(s).**
    - **Example of A × B × C interaction of $3^3$ design:**
      **ABC, $AB^2C$, $ABC^2$, $AB^2C^2$ components with 2 df each**
    - **Example of A × B interaction of $5^2$ design:**
      **AB, $AB^2$, $AB^3$, $AB^4$ components with 4 df each**
- **Defining Relationships of Interaction Components**
  - **Rule: Superscripts of interaction components are used as the weightings in the defining relationships.**
    - **Example of $AB^3C^4$ component of $5^3$ design:**
      **$X_1 + 3X_2 + 4X_3 = 0, 1, 2, 3, 4$ (Mod 5)**

Blocking and fractional factorial designs can be constructed with $2^k$, $3^k$, and $5^k$ designs. In practice, $2^k$ designs are primarily used in human factors and ergonomics research to facilitate interpretation. Each main effect and interaction in a $2^k$ factorial design has 1 df. Consequently, the entire interaction is confounded in blocking and fractional-factorial designs, whereas only components of the interaction are confounded in $3^k$ and $5^k$ factorial designs.

Orthogonal components of $3^k$ and $5^k$ designs, however, can be easily constructed by following the rule described in the center portion of this slide. Note that the superscripts of the various $3^k$ and $5^k$ components denote the weightings in the defining relationship and not the factor raised to a power. An example of using the $AB^3C^4$ component of the AxBxC interaction as a defining relationship in a $2^5$ factorial design is shown at the bottom of this slide.

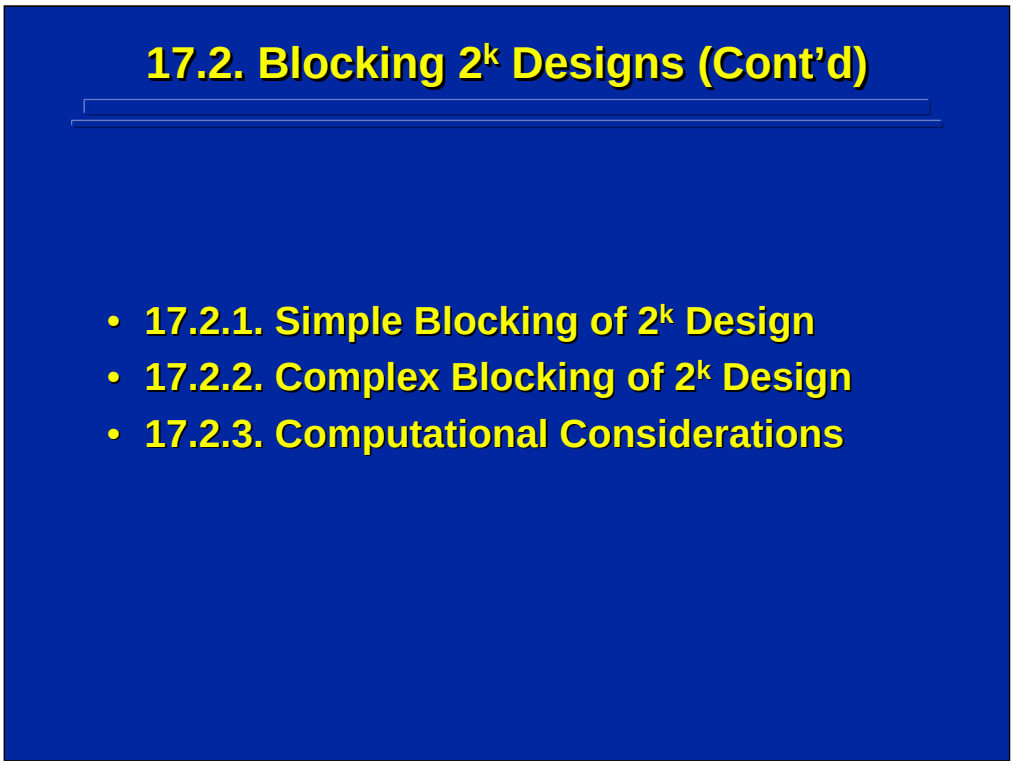



This slide shows an example of using the rule for specifying orthogonal interaction components of interactions in a $3^3$ factorial design. Each of the three main effects have 2 df. The 2 df components for the two-way interactions and the three-way interaction are shown in parenthesis on the slide. Any of these 2 df components can be used to specify a defining relationship that separates the 27 treatment conditions of the full factorial design into 3 components of 9 treatment combinations each.

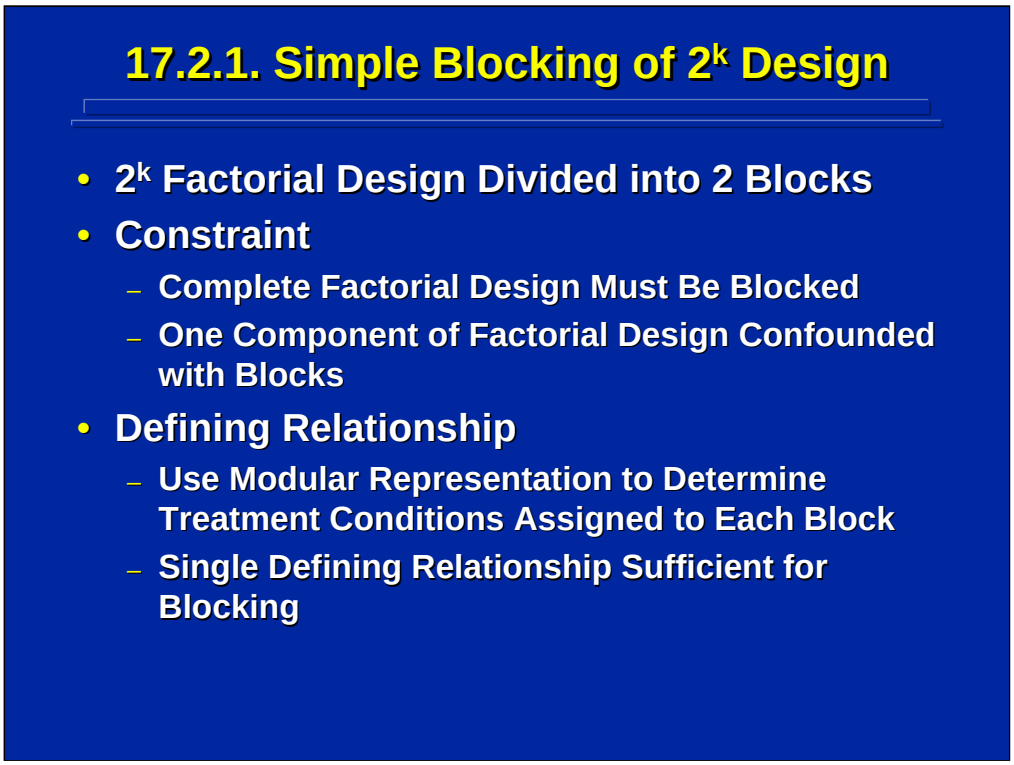



The remainder of this topic discusses the use of modular representation as a means of constructing blocks or subsets of a $2^k$ factorial design such that a nuisance variable that exists in the experiment is confounded with only a subset of the full factorial design components. For example, due to availability of experimenters it may be necessary to use more than one experimenter for data collection in a large factorial between-subjects design. Consequently, a nuisance variable due to any differences among experimenters will be confounded with a portion of the factorial design. The researcher determines the required blocking arrangement before data collection to minimize the confounding of the nuisance variable on the experiment of interest. Controlling confounding with data collection session effects in large within-subject factorial designs can also be handled through blocking procedures.

The procedures for blocking experimental designs are restricted to $2^k$ factorial designs in this topic. Since each effect in a $2^k$ factorial design has 1 df, the effect chosen as the defining relationship is totally confounded with blocks not just part of the effect. This facilitates the choice of a defining relationship. Obviously, blocking in $2^k$ factorial designs begins with two blocks and proceeds in multiples of two blocks.




Blocking of $2^k$ factorial designs is described in terms of simple and complex blocking procedures. Simple blocking of $2^k$ designs requires only one defining relationship resulting in two blocks. Complex blocking of $2^k$ designs requires more than one defining relationship yielding multiples of two blocks. This subsection ends with examples of ANOVA computations using simple and complex blocking procedures.



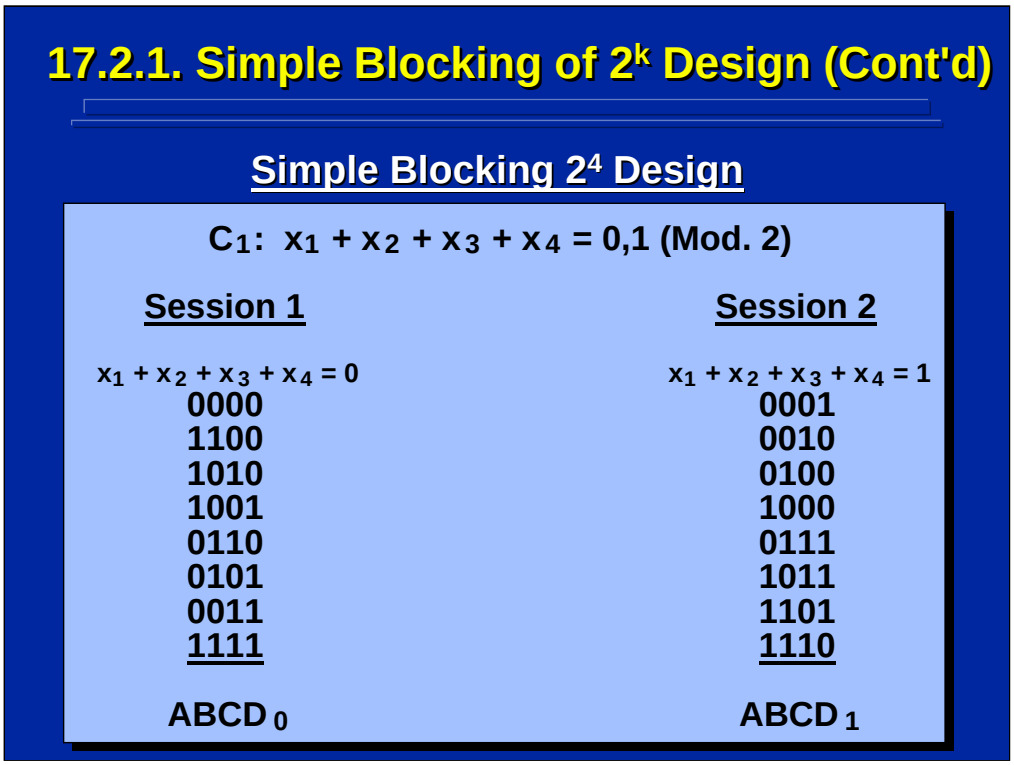

A simple blocking of a $2^k$ design divides the treatment conditions into two blocks. Any simple blocking design has the constraint that the complete factorial design is blocked such that half of the factorial design is represented in one block and the remainder is represented in the other block. Only one defining relationship is needed to construct the two blocks, and the effect chosen for the defining relationship is totally confounded with the blocking nuisance variable. Consequently, the researcher usually chooses the effect of least interest to the experiment as the defining relationship.

## 17.2.1. Simple Blocking of $2^k$ Design (Cont'd)

- **Example: $2^4$ Within-Subject Design**
  - Factors: Target Speed (A)
    Target Size (B)
    Noise Level (C)
    Display Resolution (D)
  - n = 11
  - 50 Detection Trials at Each of 16 Treatment Combinations
  - 2 Sessions of 8 Treatment Combinations
  - Confound A×B×C×D Interaction with Sessions
- **Defining Relationship**
  - $x_1 + x_2 + x_3 + x_4 = 0, 1$ (Mod. 2)

This slide describes an example of a $2^4$ within-subjects design where each of the 11 subjects receives all 16 treatment combinations in the factorial design. Since 50 target detection trials are presented at each treatment combination, each subject receives a total of 800 target detection trials to complete the experiment. The experimenter decides to divide the experiment into two sessions of 400 trials each to avoid subject fatigue.

The four-way interaction is probably of least interest to the experimenter, so it is confounded with blocks (i.e., sessions) by using the defining relationship given at the bottom of this slide. Note that careful decision must be given to choosing the defining relationship because that effect becomes confounded with blocks and cannot be evaluated separately from blocks in the subsequent data analysis.




This slide shows the eight treatment conditions of the factorial design in Mod. 2 notations that satisfy the O and 1 values, respectively, of the defining relationship ($C_1$) shown at the top for Session 1 and 2.

## 17.2.1. Simple Blocking of $2^k$ Design (Cont'd)

**ANOVA Summary Table of a Blocked, Within-Subject Design**

C $_1$: $x_1 + x_2 + x_3 + x_4 = 0, 1$ (Mod. 2)

| Source | df |
|---|---|
| Between-Subjects | |
| Subjects (S) | 10 |
| Within-Subject | |
| Blocks (AxBxCxD) | 1 |
| Blocks x S (AxBxCxDxS) | 10 |
| A | 1 |
| AxS | 10 |
| B | 1 |
| BxS | 10 |
| C | 1 |
| CxS | 10 |
| D | 1 |
| DxS | 10 |
| AxB | 1 |
| AxBxS | 10 |
| AxC | 1 |
| AxBxS | 10 |
| AxD | 1 |
| AxDxS | 10 |
| BxC | 1 |
| BxCxS | 10 |
| BxD | 1 |
| BxDxS | 10 |
| CxD | 1 |
| CxDxS | 10 |
| AxBxC | 1 |
| AxBxCxS | 10 |
| AxBxD | 1 |
| AxBxDxS | 10 |
| AxCxD | 1 |
| AxCxDxS | 10 |
| BxCxD | 1 |
| BxCxDxS | 10 |
| Total | 175 |

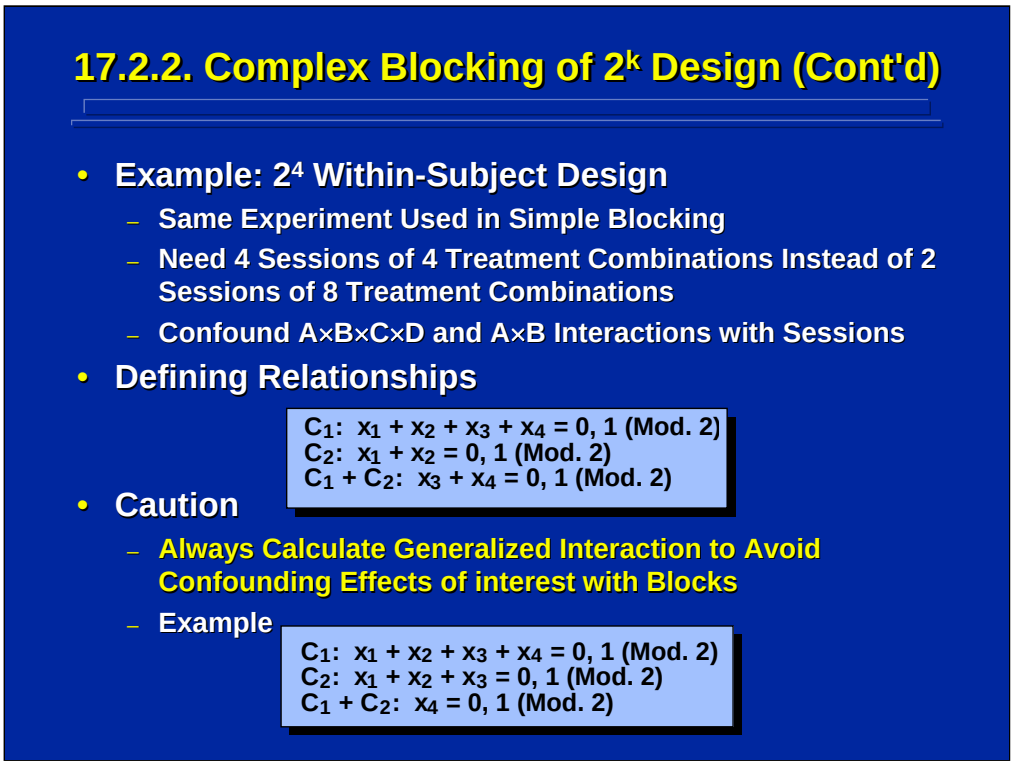
The complete ANOVA summary table for this blocked $2^4$ factorial design is shown on this slide. The Blocks main effect is Testing Session and is tested by the BlocksxSubjects interaction. Note the AxBxCxD interaction is listed in parenthesis after Blocks and the AxBxCxDxS interaction is listed in parenthesis after the BlocksxSubjects interaction because these effects are totally confounded with each other. If the Blocks effect is significant in the ANOVA, the experimenter cannot determine if the effect is due to Experimental Session (i.e. the blocking nuisance variable) or the AxBxCxD interaction due to the confounding. But, all other effects in the $2^4$ factorial design remain unconfounded with Experimental Session. Again, this underscores the notion that the effect of least interest in the experiment should be chosen as the defining relationship because there can be no unconfounded test of this effect in the blocked ANOVA design. Usually the highest-order interaction is of least interest in a factorial design and chosen for the defining relationship.

## 17.2.2. Complex Blocking of $2^k$ Design

- **Constraint**
  - Requires More Blocks Than Simple Blocking
  - Additional Blocks Confound More Effects
- **Approach**
  - Use Second Defining Relationship
  - Divide Existing Blocks into Additional Blocks
  - Determine Generalized Interactions
- **Generalized Interactions**
  - Four Blocks in $2^k$ Designs
    - $C_1$
    - $C_2$
    - $C_1 + C_2$ (Generalized Interaction)

Complex blocking is required if more than two blocks are needed for a $2^k$ design. Blocking increases in multiples of two in $2^k$ designs meaning the next levels of blocks is four, followed by eight, etc. Due to the concomitant increase in confounding effects with blocks, only four blocks are usually considered in complex blocking of $2^k$ designs. To construct these four blocks, a second defining relationship ($C_2$) must be used to divide each of the original two blocks determined by the first defining relationship ($C_1$) into two additional blocks.

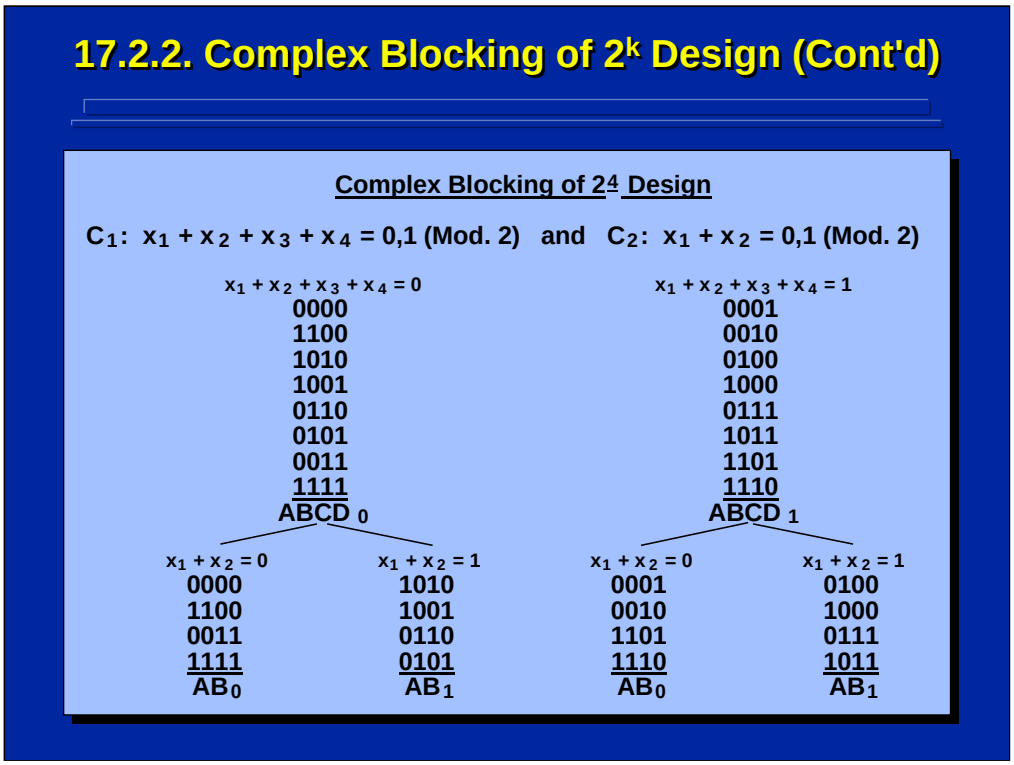
The resulting main effect of the four blocks has 3 df in the subsequent ANOVA with three 1df effects confounded with blocks. Two of these confounded effects are defined by $C_1$ and $C_2$. The third confounded effect is called the generalized interaction and is determined by adding $C_1$ and $C_2$ relationships together in Mod. 2. The resulting three confounded effects each have 1 df and represent the 3 df of the blocks main effect.




This slide extends the example of a $2^4$ within-subjects design used in simple blocking where each of the 11 subjects receives all 16 treatment combinations in the factorial design. Rather than have two blocks of eight treatment condition, four blocks of four treatment conditions are required to conduct the experiment over four sessions in this complex blocking example.

As shown in the center portion of the slide, the two defining relationships and the resulting generalized interaction were chosen such that the AxBxCxD, AxB, and CxD interactions are confounded with the four testing sessions (i.e., Blocks).

The experimenter must take care to always calculate beforehand the generalized interaction, or third confounded effect, to avoid confounding effects of interest with blocks. As shown in the bottom portion of this slide, if the highest-order interaction, AxBxCxD, and a third-order interaction, AxBxC, were chosen as the two defining relationships $C_1$ and $C_2$, then the main effect of Factor D would also be confounded with Blocks.




This slide shows the mechanics of dividing each of the two blocks from simple blocking ($C_1$) into two additional blocks of four treatment conditions in complex blocking by using $C_2$ to satisfy 0 and 1 values in Mod. 2. The four different treatment conditions of the complete $2^4$ factorial design for each of the resulting four sessions (i.e., Blocks) are listed at the bottom of this slide in Mod. 2 notation.

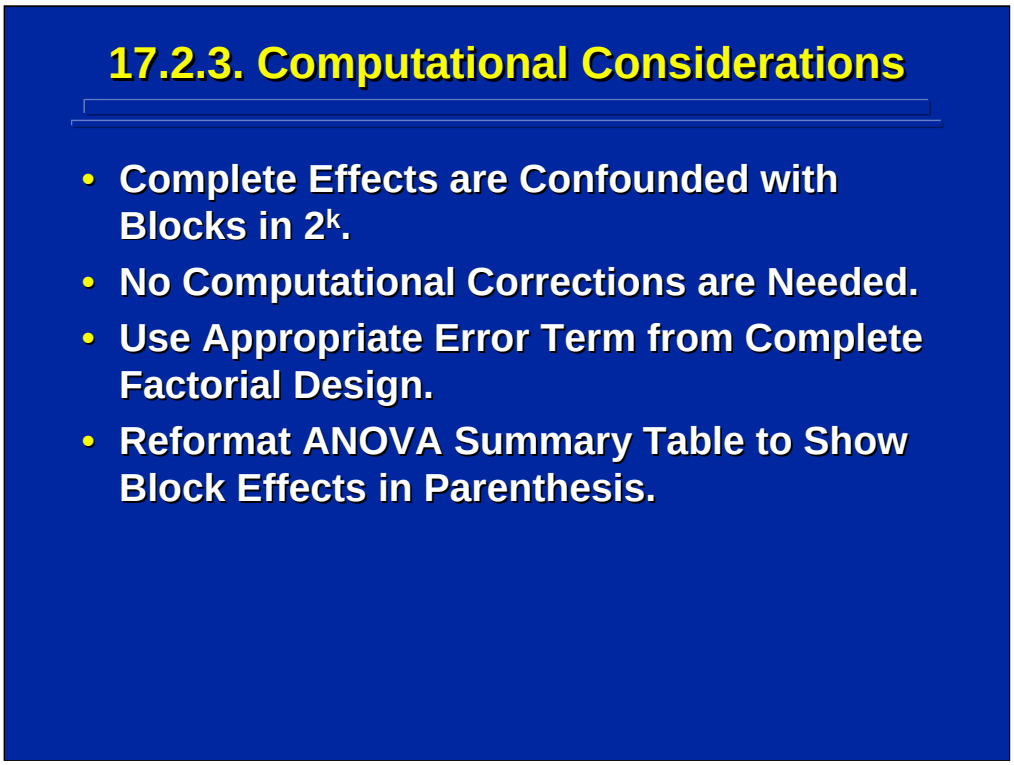
## 17.2.2. Complex Blocking of $2^k$ Design (Cont'd)

**ANOVA Summary Table of a Blocked,Within-Subject Design**

$C_1$: $x_1 + x_2 + x_3 + x_4 = 0, 1$ (Mod. 2)
$C_2$: $x_1 + x_2 = 0, 1$ (Mod. 2)
$C_1 + C_2$: $x_3 + x_4 = 0, 1$ (Mod. 2)

| Source | df |
|---|---|
| Between-Subjects | |
| Subjects (S) | 10 |
| Within-Subject | |
| Blocks (AxBxCxD, AxB, CxD) | 3 |
| Blocks x S (AxBxCxDxS, AxBxS, CxDxS) | 30 |
| A | 1 |
| AxS | 10 |
| B | 1 |
| BxS | 10 |
| C | 1 |
| CxS | 10 |
| D | 1 |
| DxS | 10 |
| AxC | 1 |
| AxCxS | 10 |
| AxD | 1 |
| AxDxS | 10 |
| BxC | 1 |
| BxCxS | 10 |
| BxD | 1 |
| BxDxS | 10 |
| AxBxC | 1 |
| AxBxCxS | 10 |
| AxBxD | 1 |
| AxBxDxS | 10 |
| AxCxD | 1 |
| AxCxDxS | 10 |
| BxCxD | 1 |
| BxCxDxS | 10 |
| Total | 175 |

The complete ANOVA summary table for a complex blocked $2^4$ factorial design is shown on this slide. The two defining relationships and the generalized interaction are stated at the top of this slide. These three $C_1$, $C_2$, and $C_1 + C_2$ effects are the three interactions confounded with blocks as shown in parenthesis after blocks. The interaction of each of these three effects and subjects are confounded with BlocksxSubjects and are listed in parenthesis.




Since complete effects in $2^k$ factorial designs are confounded with blocks, standard statistical analysis packages can be used to conduct the ANOVA without computational correction on any effect of interest. The SS of the effects confounded with blocks are merely added together to determine the Blocks SS, and the ANOVA Summary Table is reformatted accordingly. The same error terms used for between-subjects and within-subjects ANOVA designs are used for the blocked designs.

## 17.2.3. Computational Considerations (Cont'd)

**Between-Subjects Design**

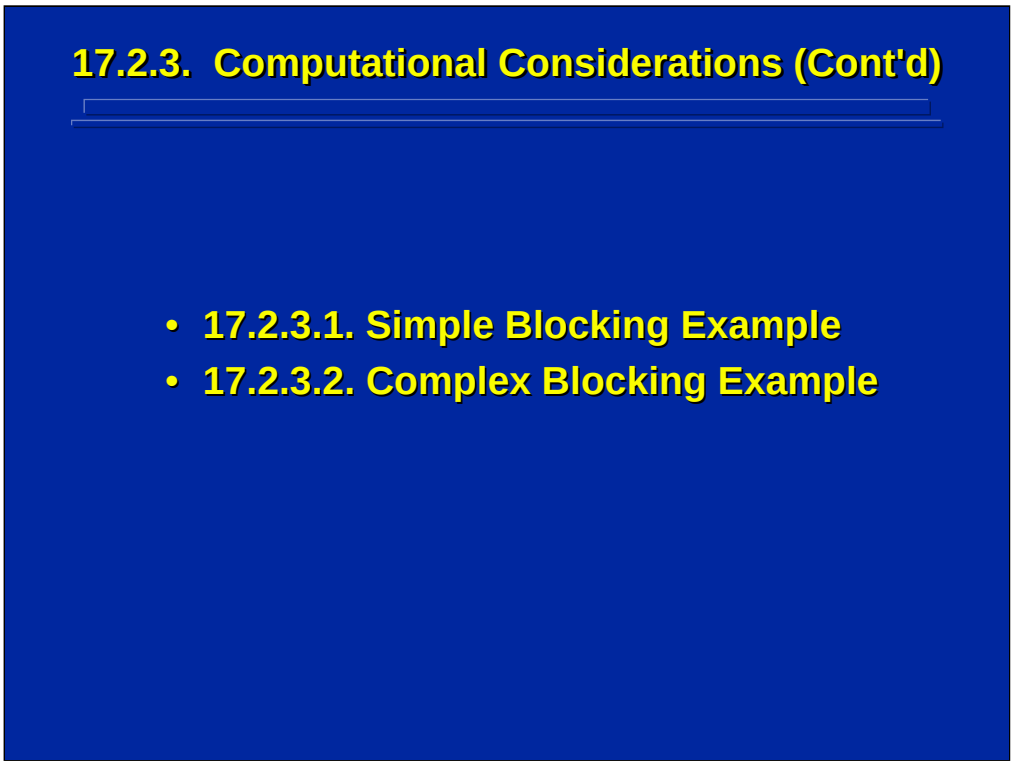
| Source | df |
|---|---|
| **Between-Subjects** | |
| Blocks (AxBxC) | (g-1) |
| A | (a-1) |
| B | (b-1) |
| C | (c-1) |
| AxB | (a-1)(b-1) |
| AxC | (a-1)(c-1) |
| BxC | (b-1)(c-1) |
| S/ABC | abc(n-1) |
| Total | abcn-1 |

**Within-Subject Design**

| Source | df |
|---|---|
| **Between-Subjects** | |
| Subjects (S) | (n-1) |
| **Within-Subject** | |
| Blocks (AxBxC) | (g-1) |
| Blocks x S (AxBxCxS) | (g-1)(n-1) |
| A | (a-1) |
| AxS | (a-1)(n-1) |
| B | (b-1) |
| BxS | (b-1)(n-1) |
| C | (c-1) |
| CxS | (c-1)(n-1) |
| AxB | (a-1)(b-1) |
| AxBxS | (a-1)(b-1)(n-1) |
| AxC | (a-1)(c-1) |
| AxCxS | (a-1)(c-1)(n-1) |
| BxC | (b-1)(c-1) |
| BxCxS | (b-1)(c-1)(n-1) |
| Total | abcn-1 |

By way of example, this slide shows a comparison of a between-subjects and a within-subjects $2^3$ factorial design that is conducted in two blocks using the AxBxC interaction as the defining relationship for simple blocking. Note that both designs show the AxBxC interaction confounded with blocks, and calculation of this interaction effect can be used to calculate the Block SS.

The error used to test the Block effect, however, differs depending on the type of design. The between-subjects design uses the usual S/ABC effect as the error term for effects including Blocks, and the within-subjects design alternative uses the interaction with blocks (i.e., AxBxCxS) as the error term for testing Blocks.




Two examples are provided in this subsection to demonstrate the ANOVA computations of within-subjects simple and complex blocking designs. The Slater and Williges (2006) appendix provides the SAS analysis procedures for conducting the subsequent ANOVA on each example problem.

## 17.2.3.1. Simple Blocking Example

- **Example Problem: Testing was conducted on a new computerized target detection system. The detection system evaluates four different dimensions (i.e., target speed, target size, noise level, and display resolution) each with two settings. Five soldiers have been recruited to participate in the testing of the new system. For each of the 16 dimension combinations, 100 detection trials per soldier are completed and a percentage is computed. Because of the number of trials (1600 trials per soldier), the testing procedure is too lengthy to complete in one day, so it will be conducted in two sessions over two days. Do the settings have an effect on the percentage of targets detected? ($p < 0.01$) Also, is there an effect due to the blocking of the data collection into two sessions?**

(Click in this red rectangle to see SAS calculations for this example.)

This example describes the same $2^4$ within-subjects target detection problem used in simple blocking, but only uses a sample size of five. The blocking requirement is to divide the data collection of the full factorial design into two equal sessions consisting of eight different treatment conditions of 100 trials each.

## 17.2.3.1. Simple Blocking Example (Cont'd)

Target Speed (A), Target Size (B), Noise Level (C), Display Resolution (D)

$C_1$: $x_1 + x_2 + x_3 + x_4 = 0,1$ (Mod. 2)

| Session 1 | Session 2 |
|---|---|
| $x_1 + x_2 + x_3 + x_4 = 0$ | $x_1 + x_2 + x_3 + x_4 = 1$ |
| 0000 | 0001 |
| 1100 | 0010 |
| 1010 | 0100 |
| 1001 | 1000 |
| 0110 | 0111 |
| 0101 | 1011 |
| 0011 | 1101 |
| 1111 | 1110 |
| $ABCD_0$ | $ABCD_1$ |

(Click in this red rectangle to see SAS calculations for this example.)

The fourth-order interaction (AxBxCxD) is used as the defining relation, $C_1$. This slide shows the resulting eight treatment combinations in Mod. 2 notation for each session.

## 17.2.3.1. Simple Blocking Example (Cont'd)

**Target Speed (A), Target Size (B), Noise Level (C), Display Resolution (D)**

| Session 1 ($ABCD_0$) | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0000 | 1100 | 1010 | 1001 | 0110 | 0101 | 0011 | 1111 |
| 0.50 | 0.09 | 0.75 | 0.28 | 0.02 | 0.70 | 0.78 | 0.14 |
| 0.23 | 0.23 | 0.48 | 0.22 | 0.43 | 0.67 | 0.89 | 0.27 |
| 0.45 | 0.14 | 0.38 | 0.39 | 0.14 | 0.90 | 0.64 | 0.08 |
| 0.66 | 0.37 | 0.89 | 0.08 | 0.27 | 0.87 | 0.50 | 0.31 |
| 0.37 | 0.46 | 0.66 | 0.44 | 0.19 | 0.76 | 0.40 | 0.25 |

| Session 2 ($ABCD_1$) | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0001 | 0010 | 0100 | 1000 | 0111 | 1011 | 1101 | 1110 |
| 0.11 | 0.05 | 0.32 | 0.74 | 0.99 | 0.05 | 0.31 | 0.99 |
| 0.77 | 0.60 | 0.41 | 0.55 | 0.68 | 0.39 | 0.59 | 0.81 |
| 0.27 | 0.16 | 0.33 | 0.43 | 0.68 | 0.40 | 0.71 | 0.77 |
| 0.33 | 0.21 | 0.11 | 0.67 | 0.41 | 0.62 | 0.61 | 0.59 |
| 0.41 | 0.10 | 0.56 | 0.77 | 0.77 | 0.57 | 0.59 | 0.54 |

(Click in this red rectangle to see SAS calculations for this example.)

The set of probability of detection scores across 100 trials for each soldier in the full within-subjects, factorial design is shown on this slide. The eight conditions each soldier experienced in Session 1 are shown on the top portion in Mod. 2 notation, and the eight conditions each soldier experienced in Session 2 are shown in the bottom portion in Mod. 2 notation. This division of the 16 treatment combinations follows the blocking represented in the previous slide.

## 17.2.3.1. Simple Blocking Example (Cont'd)

ANOVA Summary Table

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Between-Subjects | | | | |
| Subjects (S) | 4 | 0.0672 | 0.0168 | |
| Within-Subject | | | | |
| Session (AxBxCxD) | 1 | 0.1232 | 0.1232 | 4.75 |
| SessionxS (AxBxCxDxS) | 4 | 0.1038 | 0.0259 | |
| Target Speed (A) | 1 | 0.0022 | 0.0022 | 0.05 |
| AxS | 4 | 0.1608 | 0.0402 | |
| Target Size (B) | 1 | 0.0022 | 0.0022 | 0.20 |
| BxS | 4 | 0.0451 | 0.0113 | |
| Noise Level (C) | 1 | 0.0101 | 0.0101 | 0.22 |
| CxS | 4 | 0.1814 | 0.0453 | |
| Display Resolution (D) | 1 | 0.1022 | 0.1022 | 6.96 |
| DxS | 4 | 0.0588 | 0.0147 | |
| AxB | 1 | 0.1232 | 0.1232 | 2.65 |
| AxBxS | 4 | 0.1857 | 0.0464 | |
| AxC | 1 | 0.0806 | 0.0806 | 4.57 |
| AxCxS | 4 | 0.7059 | 0.0176 | |
| AxD | 1 | 1.2450 | 1.2450 | 56.33* |
| AxDxS | 4 | 0.0884 | 0.0221 | |
| BxC | 1 | 0.0361 | 0.0361 | 3.52 |
| BxCxS | 4 | 0.0410 | 0.0103 | |
| BxD | 1 | 0.2184 | 0.2184 | 7.80 |
| BxDxS | 4 | 0.1119 | 0.0275 | |
| CxD | 1 | 0.0018 | 0.0018 | 0.05 |
| CxDxS | 4 | 0.1323 | 0.0331 | |
| AxBxC | 1 | 0.0092 | 0.0092 | 0.34 |
| AxBxCxS | 4 | 0.1102 | 0.0275 | |
| AxBxD | 1 | 0.0312 | 0.0312 | 1.29 |
| AxBxDxS | 4 | 0.0971 | 0.0243 | |
| AxCxD | 1 | 0.4234 | 0.4234 | 6.69 |
| AxCxDxS | 4 | 0.2532 | 0.0633 | |
| BxCxD | 1 | 0.6734 | 0.6734 | 21.89* |
| BxCxDxS | 4 | 0.1231 | 0.0308 | |
| Total | 79 | 4.9133 | | |

*p < 0.01

(Click in this red rectangle to see SAS calculations for this example.)

The ANOVA on the complete $2^4$ within-subjects design can be conducted on this data set using standard procedures. The session effect is merely the AxBxCxD interaction and is tested by the AxBxCxDxS interaction.

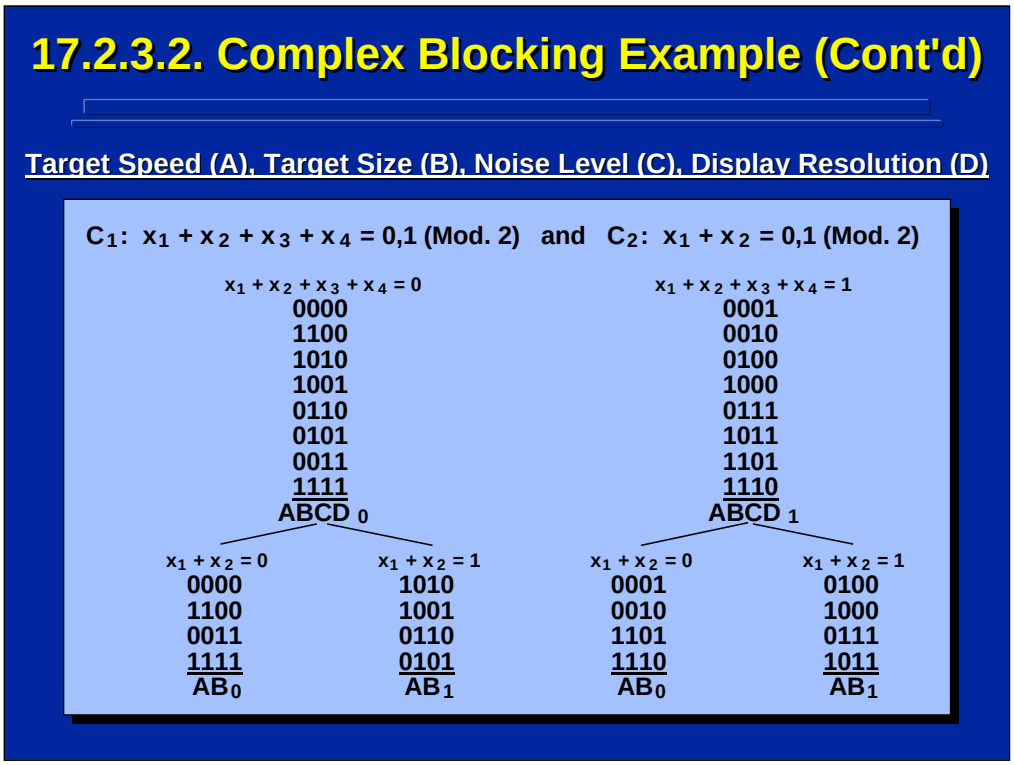
The complete ANOVA computations for this simple blocking problem using SAS is presented in the Slater and Williges (2006) appendix, and the resulting ANOVA Summary Table is shown on this slide. The abbreviations for the four independent variables are listed as A, B, C, and D to make them compatible with the previous slides for this example. Normally, a meaningful abbreviation is chosen for each factor.

## 17.2.3.2. Complex Blocking Example

- **Example Problem: Testing was conducted on a new computerized target detection system. The detection system evaluates four different dimensions (i.e., target speed, target size, noise level, and display resolution) each with two settings. Five soldiers have been recruited to participate in the testing of the new system. For each of the 16 dimension combinations, 100 detection trials per soldier are completed and a percentage is computed. Because of the number of trials (1600 trials per soldier), the testing procedure is too lengthy to complete in one day, so it will be conducted in four sessions over four days. Do the settings have an effect on the percentage of targets detected? ($p < 0.01$) Also, is there an effect due to the blocking of the data collection into four sessions?**

(Click in this red rectangle to see SAS calculations for this example.)

This example describes the same $2^4$ within-subjects target detection problem used in complex blocking, but only uses a sample size of five. The blocking requirement is to divide the data collection of the full factorial design into four equal sessions consisting of four different treatment conditions of 100 trials each.




(Click in this red rectangle to see SAS calculations for this example.)

The fourth-order interaction (AxBxCxD) is used as the first defining relationship, $C_1$, and a two-way interaction (AxB) is used as the second defining relationship, $C_2$. Consequently, the generalized interaction (i.e., $C_1$ + $C_2$) is CxD. This slide shows the resulting four treatment combinations in Mod. 2 notation for each of the four data collection sessions.

## 17.2.3.2. Complex Blocking Example (Cont'd)

**Target Speed (A), Target Size (B), Noise Level (C), Display Resolution (D)**

| Session 1 ($ABCD_0$ and $AB_0$) | | | |
|---|---|---|---|
| 0000 | 1100 | 0011 | 1111 |
| 0.50 | 0.09 | 0.78 | 0.14 |
| 0.23 | 0.23 | 0.89 | 0.27 |
| 0.45 | 0.14 | 0.64 | 0.08 |
| 0.66 | 0.37 | 0.50 | 0.31 |
| 0.37 | 0.46 | 0.40 | 0.25 |

| Session 2 ($ABCD_0$ and $AB_1$) | | | |
|---|---|---|---|
| 1010 | 1001 | 0110 | 0101 |
| 0.75 | 0.28 | 0.02 | 0.70 |
| 0.48 | 0.22 | 0.43 | 0.67 |
| 0.38 | 0.39 | 0.14 | 0.90 |
| 0.89 | 0.08 | 0.27 | 0.87 |
| 0.66 | 0.44 | 0.19 | 0.76 |

| Session 3 ($ABCD_1$ and $AB_0$) | | | |
|---|---|---|---|
| 0001 | 0010 | 1101 | 1110 |
| 0.11 | 0.05 | 0.31 | 0.99 |
| 0.77 | 0.60 | 0.59 | 0.81 |
| 0.27 | 0.16 | 0.71 | 0.77 |
| 0.33 | 0.21 | 0.61 | 0.59 |
| 0.41 | 0.10 | 0.59 | 0.54 |

| Session 4 ($ABCD_1$ and $AB_1$) | | | |
|---|---|---|---|
| 0100 | 1000 | 0111 | 1011 |
| 0.32 | 0.74 | 0.99 | 0.05 |
| 0.41 | 0.55 | 0.68 | 0.39 |
| 0.33 | 0.43 | 0.68 | 0.40 |
| 0.11 | 0.67 | 0.41 | 0.62 |
| 0.56 | 0.77 | 0.77 | 0.57 |

(Click in this red rectangle to see SAS calculations for this example.)

The complete set of probability of detection scores across 100 trials for each soldier in the full within-subjects, factorial design is shown on this slide. The data are grouped by the four conditions in Mod. 2 notation that each soldier received in each of the four sessions. This division of the 16 treatment combinations follows the blocking represented in the previous slide.

## 17.2.3.2. Complex Blocking Example (Cont'd)

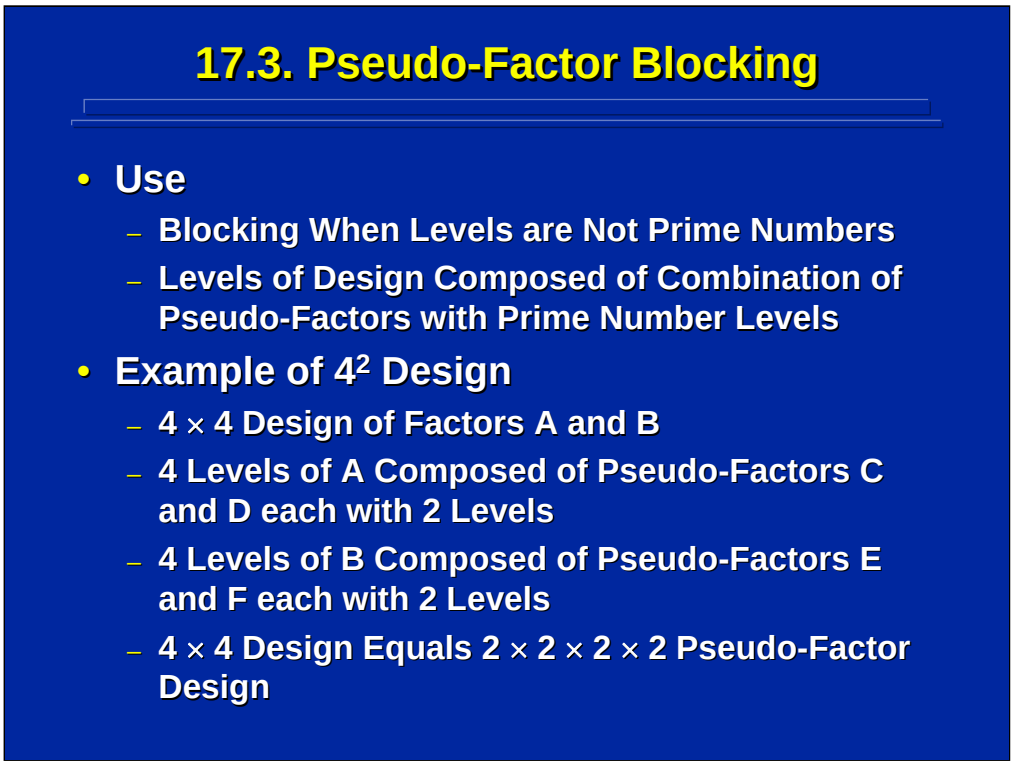
ANOVA Summary Table

| Source | df | SS | MS | F |
|---|---|---|---|---|
| **Between-Subjects** | | | | |
| Subjects (S) | 4 | 0.0672 | 0.0168 | |
| **Within-Subject** | | | | |
| Session (AxBxCxD, AxB, CxD) | 3 | 0.2483 | 0.0828 | 2.35 |
| SessionxS (AxBxCxDxS, AxBxS, CxDxS) | 12 | 0.4217 | 0.0351 | |
| Target Speed (A) | 1 | 0.0022 | 0.0022 | 0.05 |
| AxS | 4 | 0.1608 | 0.0402 | |
| Target Size (B) | 1 | 0.0022 | 0.0022 | 0.20 |
| BxS | 4 | 0.0451 | 0.0113 | |
| Noise Level (C) | 1 | 0.0101 | 0.0101 | 0.22 |
| CxS | 4 | 0.1814 | 0.0454 | |
| Display Resolution (D) | 1 | 0.1022 | 0.1022 | 6.96 |
| DxS | 4 | 0.0588 | 0.0147 | |
| AxC | 1 | 0.0806 | 0.0806 | 4.57 |
| AxCxS | 4 | 0.0706 | 0.0176 | |
| AxD | 1 | 1.2450 | 1.2450 | 56.33* |
| AxDxS | 4 | 0.0884 | 0.0221 | |
| BxC | 1 | 0.0361 | 0.0361 | 3.52 |
| BxCxS | 4 | 0.0410 | 0.0103 | |
| BxD | 1 | 0.2184 | 0.2184 | 7.80 |
| BxDxS | 4 | 0.1119 | 0.0278 | |
| AxBxC | 1 | 0.0092 | 0.0092 | 0.34 |
| AxBxCxS | 4 | 0.1102 | 0.0275 | |
| AxBxD | 1 | 0.0312 | 0.0312 | 1.29 |
| AxBxDxS | 4 | 0.0971 | 0.0242 | |
| AxCxD | 1 | 0.4234 | 0.4234 | 6.69 |
| AxCxDxS | 4 | 0.2532 | 0.0633 | |
| BxCxD | 1 | 0.6734 | 0.6734 | 21.89* |
| BxCxDxS | 4 | 0.1231 | 0.0308 | |
| Total | 79 | 4.9133 | | |

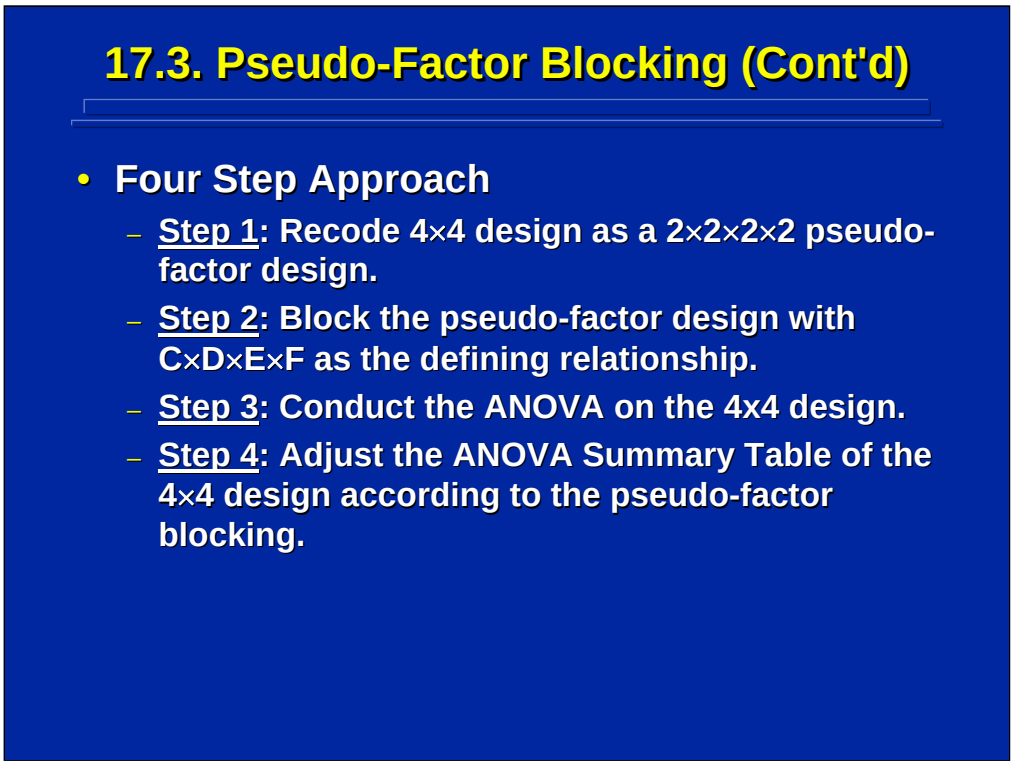
*p < 0.01

(Click in this red rectangle to see SAS calculations for this example.)

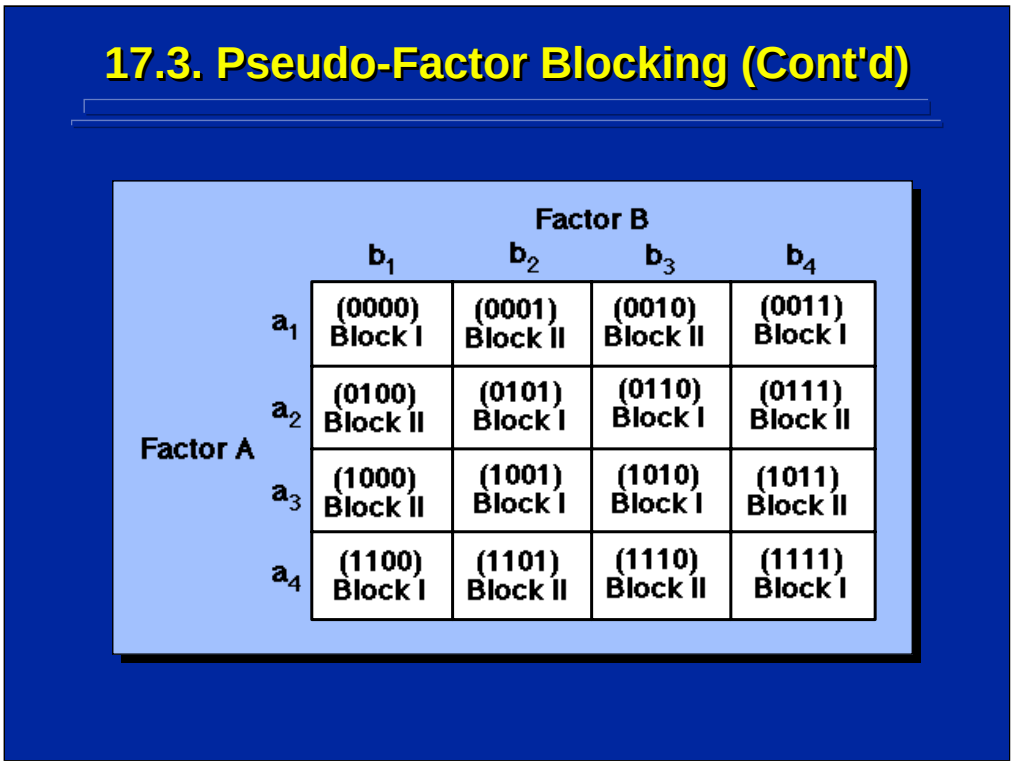
The ANOVA on the complete $2^4$ within-subjects design can be conducted on this data set using standard procedures. The session main effect with 3df can be calculated separately instead of adding together the AxBxCxD, AxB, and CxD interactions that are confounded with sessions. Likewise, the SessionsxS interaction is used as the error term for testing Sessions and can be calculated with 12df instead of the AxBxCxDxS, AxBxS, and CxDxS interactions in this within-subjects design. The complete ANOVA computations for this simple blocking problem using SAS is presented in the Slater and Williges (2006) appendix, and the resulting ANOVA Summary Table is shown on this slide. The abbreviations for the four independent variables are listed as A, B, C, and D to make them compatible with the previous slides for this example. Normally, a meaningful abbreviation is chosen for each factor.



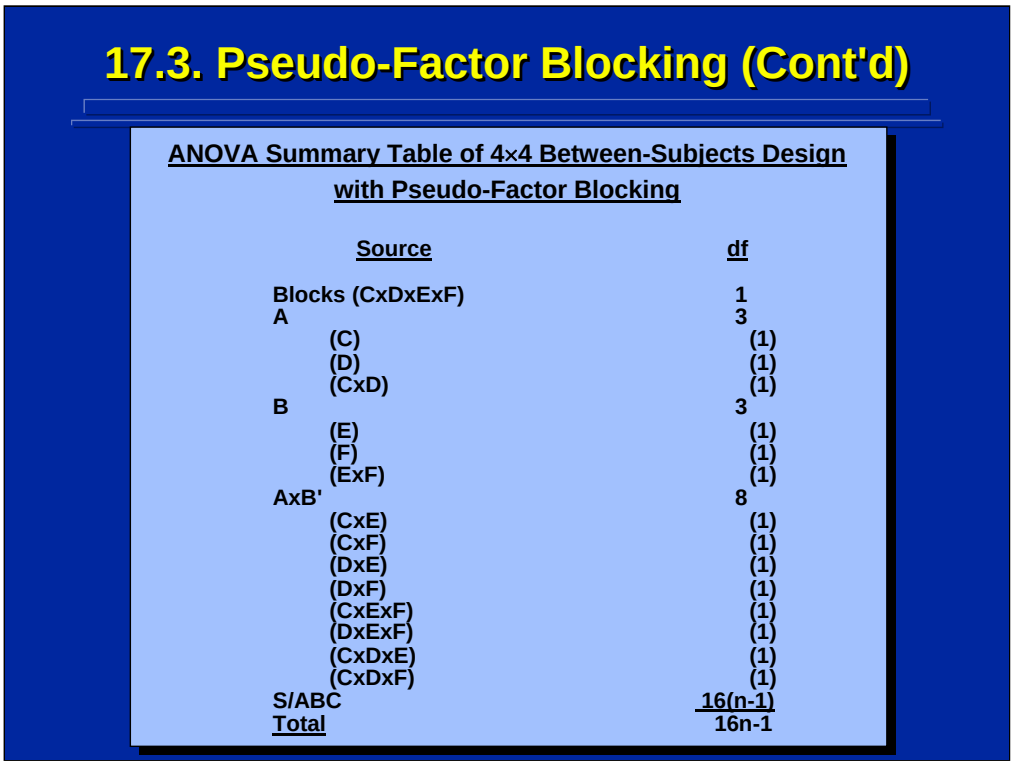

Pseudo-factors can be used to construct blocks in certain situations when the levels of the actual design are composed of a combination of factors consisting of prime number levels. The factors in this combination are called pseudo-factors. An example of a 4x4 factorial design with factors A and B is shown on this slide. These 16 treatment combinations can be blocked as a $2^4$ factorial design in which Factor A is designated as pseudo-factors C and D with 2 levels each, and Factor B is designated as pseudo-factors E and F with two factors each.




This slide summarizes a four-step approach for blocking the 4x4 design into two blocks of eight treatment conditions. First, the design is recoded as pseudo-factors. Second, the pseudo-factors are blocked using the highest-order interaction as the defining relationship. Third, the ANOVA is conducted on the 4x4 design. And, fourth, the ANOVA Summary Table is adjusted by using the $2^4$ pseudo-factor blocking design.

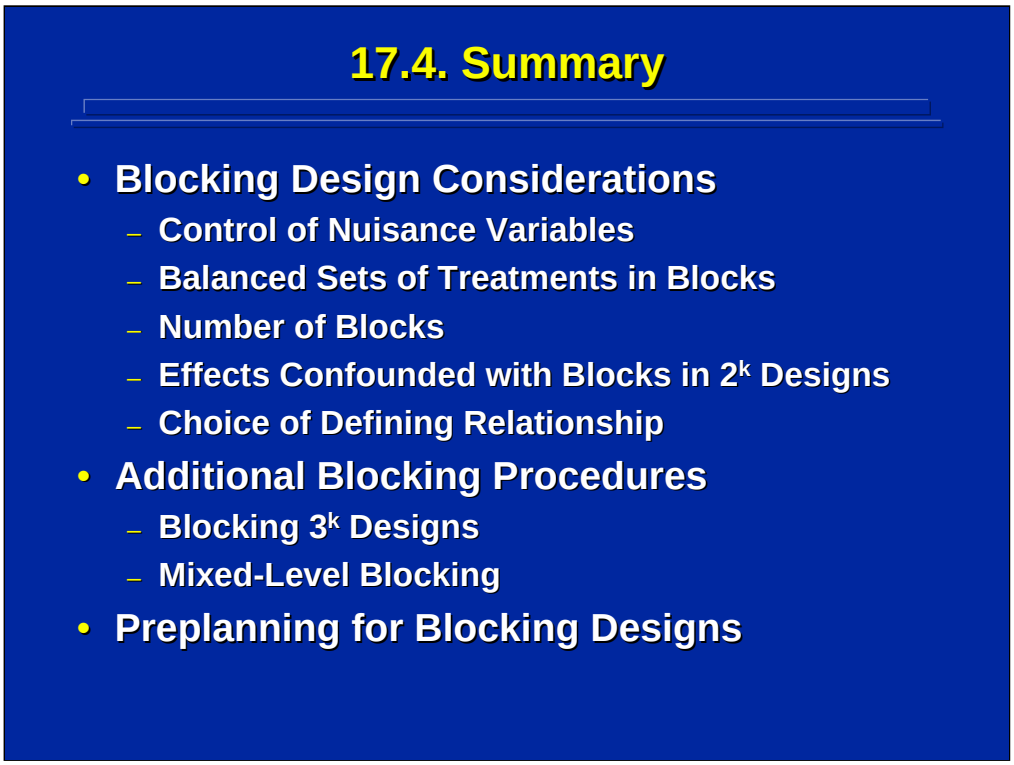



The design matrix shown on this slide is the result of the first two steps in pseudo-factor blocking of the original 4x4 factorial design shown as the rows and columns of the 16 cell design matrix. The C, D, E, and F pseudo-factor designation is shown in Mod. 2 notation for each of the 16 cells in the design matrix. By using the CxDxExF interaction of the pseudo-factors as the defining relationship, the original 4x4 factorial design is divided into two blocks of 8 treatment combinations each according to the designation shown in this slide.



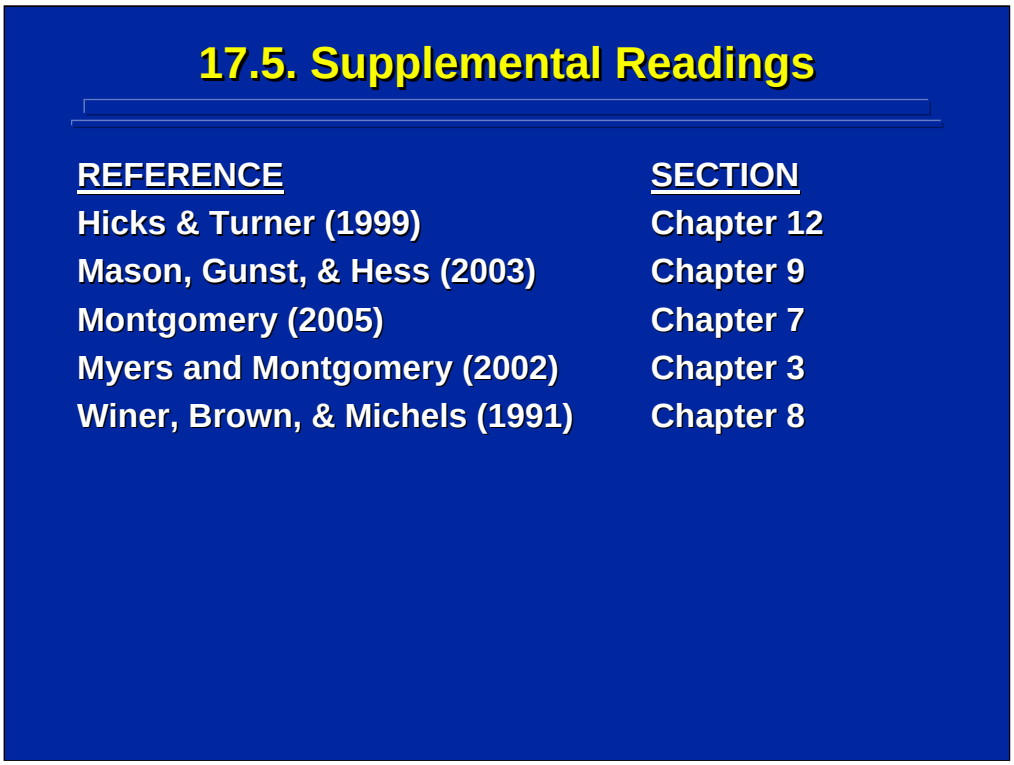

In Steps 3 and 4 of the pseudo-factor blocking process, the ANOVA is conducted on the 4x4 factorial design, and the final ANOVA Summary Table is restated according to the pseudo-factor blocking. The 4x4 ANOVA Summary Table shown on this slide shows the pseudo-factor relationship in parenthesis for the 4x4 between-subjects experimental design. Note that Blocks is confounded with the four-way interaction of the pseudo-factors.

The main effects of Factors A and B each have 3 df that represent the main effects of the two pseudo-factors and their interaction. The AxB interaction has only 8 df rather than 9 df since 1 df is confounded with blocks. Consequently, it is listed as AxB' to designate an incomplete interaction effect. Note that the main effects of A and B are unconfounded and only 1 df of the AxB interaction is confounded with blocks. The key is to choose only one pseudo-factor effect that is confounded with the AxB interaction as the defining relationship in order to minimize the blocking effect on the original 4x4 factorial design and keep the A and B effects unconfounded when constructing blocks using pseudo-factors.




Blocking in ANOVA is used when a nuisance variable such as data collection days is not crossed with the factorial design and causes confounding. Experimental constraints and the choice of the factorial design determine the number of blocks that can be used. Modular representation facilitates determination of balanced sets of treatments within blocks. These balanced treatments represent entire main effects and interactions in $2^k$ factorial designs, and the experimenter should consider using this type of design if blocking a nuisance variable is required. Defining relationships specify which effects in the factorial design are confounded with blocks and should be considered carefully to avoid confounding effects of interest with the nuisance variable.

Incomplete blocking procedures can be extended to pseudo-blocking and mixed-level blocking. Blocking can be extended to $3^k$ designs by using Mod. 3 notation if the experimenter is willing to confound components of effects. Mixed-level blocking can be used in designs like a 2x2x2x3 factorial design. If the three-level factor cannot be reduced to two levels to make a $2^4$ design, the design is a $2^3$x3 design. The experimenter can block the $2^3$ portion and then cross it with the three-level factor. Winer et al. (1991, pp. 647-660) provide a complete description of $3^k$ blocking and mixed-level blocking procedures. Careful preplanning on the part of the experimenter is needed to design successful ANOVA experiments that block the nuisance variable appropriately.




All these texts provide a discussion of blocking designs used in ANOVA. Winer, et al. (1991) provide a detailed description of modular arithmetic and the modular representation used in this topic to construct simple and complex blocking designs.

# Topic 18. Fractional-Factorial ANOVA Designs


18.1. $2^{k-p}$ Fractional Replicates
- 18.1.1. Design Construction
- 18.1.2. Computational Considerations
- 18.1.3. Design Resolution

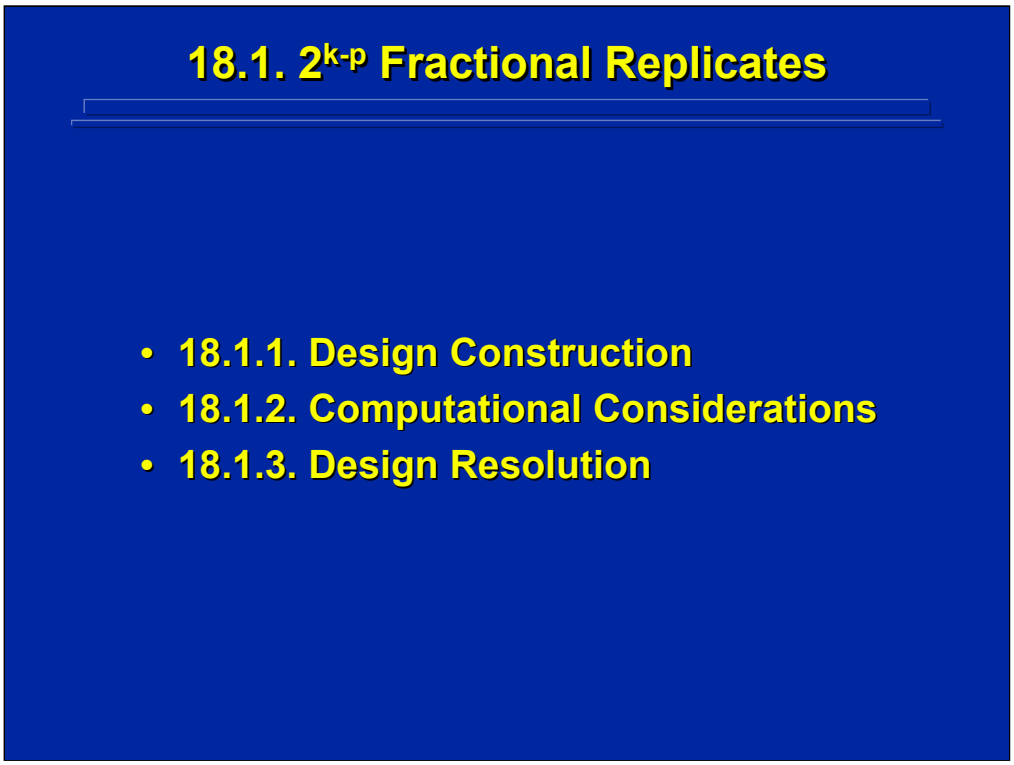
18.2. Latin Square ANOVA Designs
- 18.2.1. Design Construction
- 18.2.2. Computational Considerations
- 18.2.3. Design Constraints

18.3. Summary

18.4. Supplemental Readings


When large-scale factorial designs are used in human factors and ergonomics research, a variety of time, money, and equipment availability constraints may make the complete factorial design unfeasible and data collection must be restricted to a portion of the design. Obviously, some effects in the full factorial design cannot be evaluated, and some effects will be confounded with others if data are not collected on the entire design. The experimenter must select a subset of the complete design that probably will yield the most useful data.

Fractional-factorial designs are ANOVA designs in which only a fractional portion of the complete factorial design is observed. Fractional replications of $2^k$, $3^k$, and $5^k$ designs can be constructed through modular representation. Most often $2^{k-p}$ fractional replicates are used in human factors research to avoid subsequent confounding of partial effects present in $3^k$ and $5^k$ designs. When the experimenter is only interested in testing the main effects of three factors of interest, a special category of fractional-factorial designs called Latin square designs can be used to specify the subset of treatment conditions to observe. Consequently, this topic describes both $2^{k-p}$ fractional replicates and Latin square designs. A summary of these procedures and suggested additional readings are provided at the end of this topic.



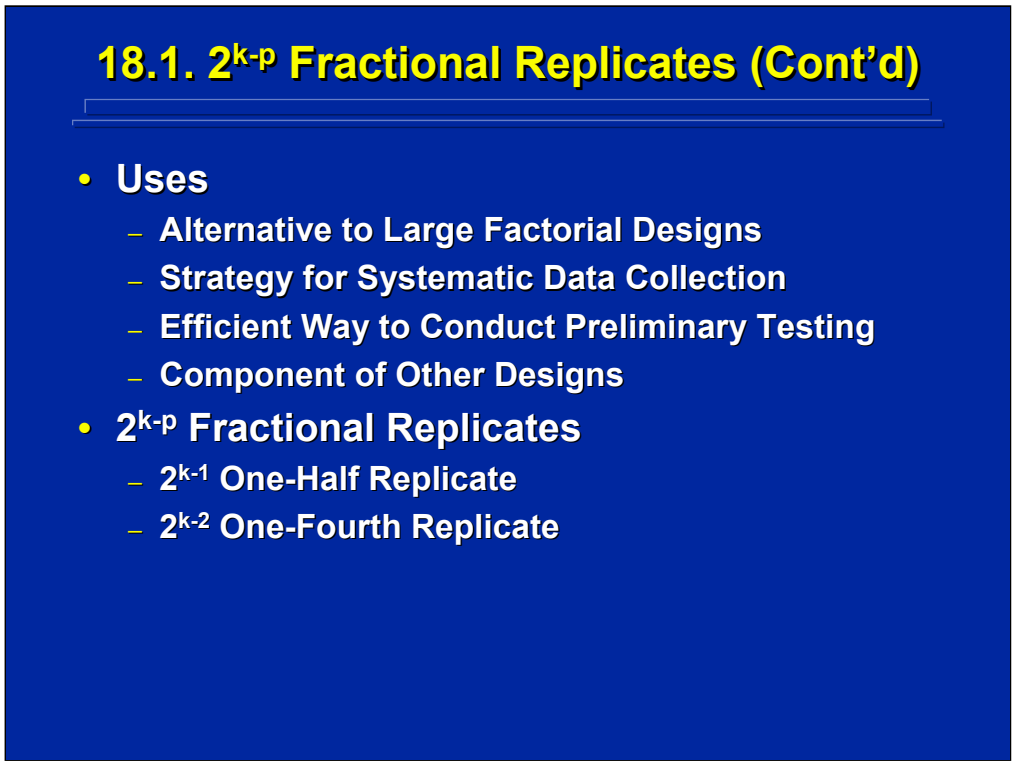

Fractional replicates of $2^k$ ANOVA designs are referred to as $2^{k-p}$ designs. This subsection describes the construction of these designs using Mod. 2 representation as described in Chapter 9 by Winer, et al. (1991), computational procedures for conducting an ANOVA on the results obtained from $2^{k-p}$ designs, and the concept of design resolution which is important in selecting the appropriate $2^{k-p}$ fractional replicate.

## 18.1. $2^{k-p}$ Fractional Replicates (Cont'd)

- **Description**
  - **Fractional Replicate: Subset of Complete Factorial Design Chosen by Modular Representation**
  - **Identity Relationship (I): Effect Used to Construct the Fractional Replicate and Held Constant**
  - **Aliases: Effects Totally Confounded in Fractional Replicate**
- **Assumption**
  - **Interaction Aliases Do Not Exist or Are Negligible**
- **Information Impact**
  - **Loss of Identity Relationship Effect**
  - **Confounded Effects in Alias Structure**

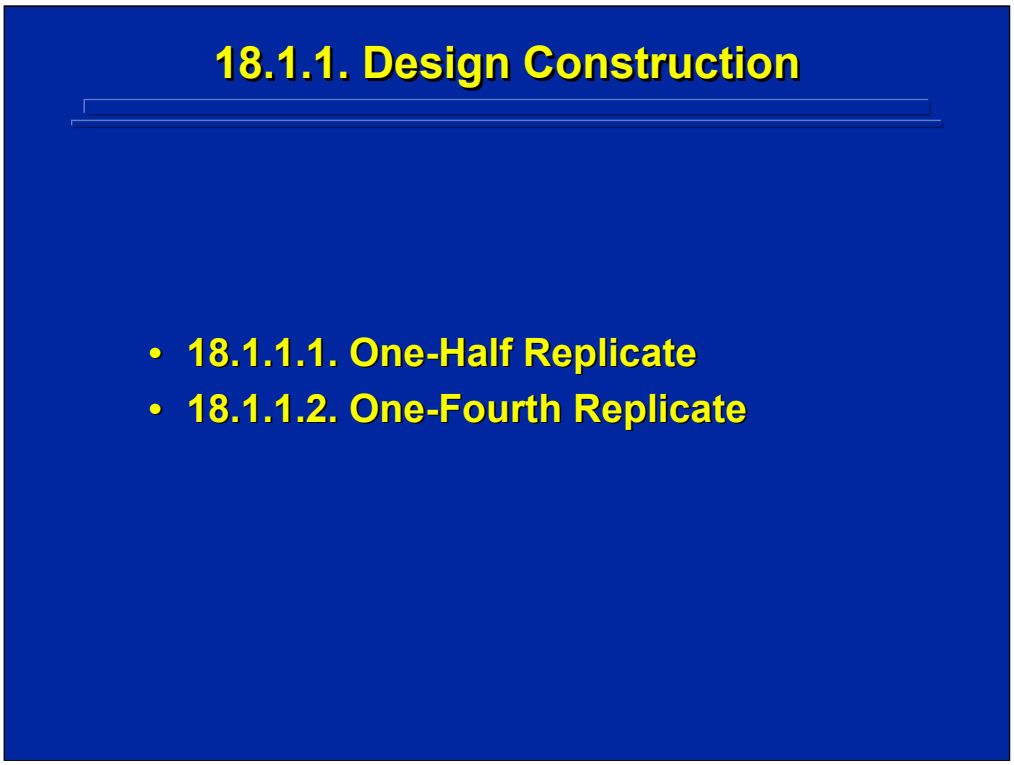
There are three characteristics of fractional replicates. First, they represent a subset of the complete factorial design. The same mechanics of modular representation used in blocking designs are used in fractional replicates, but only one block is chosen for data collection in the fractional replicate. Second, the identity relationship used to form the fractional replicate is the value of the defining relationship used for the chosen block. The identity relationship effect cannot be evaluated in the experiment because only one level of this effect is evaluated in the experiment. And, third, since data from only a fractional portion of the total design are collected, some of the effects in the complete factorial design are confounded with each other. The effects confounded with the effects being tested in the fractional replicate are referred to as aliases. Effects listed as aliases are assumed not to exist or to be negligible in order to interpret the effects of interest.

Since the complete factorial design is not used, the information impact needs to be considered during experimental design. The experimenter must pay careful attention to choosing the identity relationship and the resulting alias structure in constructing a fractional replicate in order to obtain the maximum research benefit from the fractional-factorial experiment since the effect in the alias structure cannot be evaluated and the effects in the alias structure are confounded.

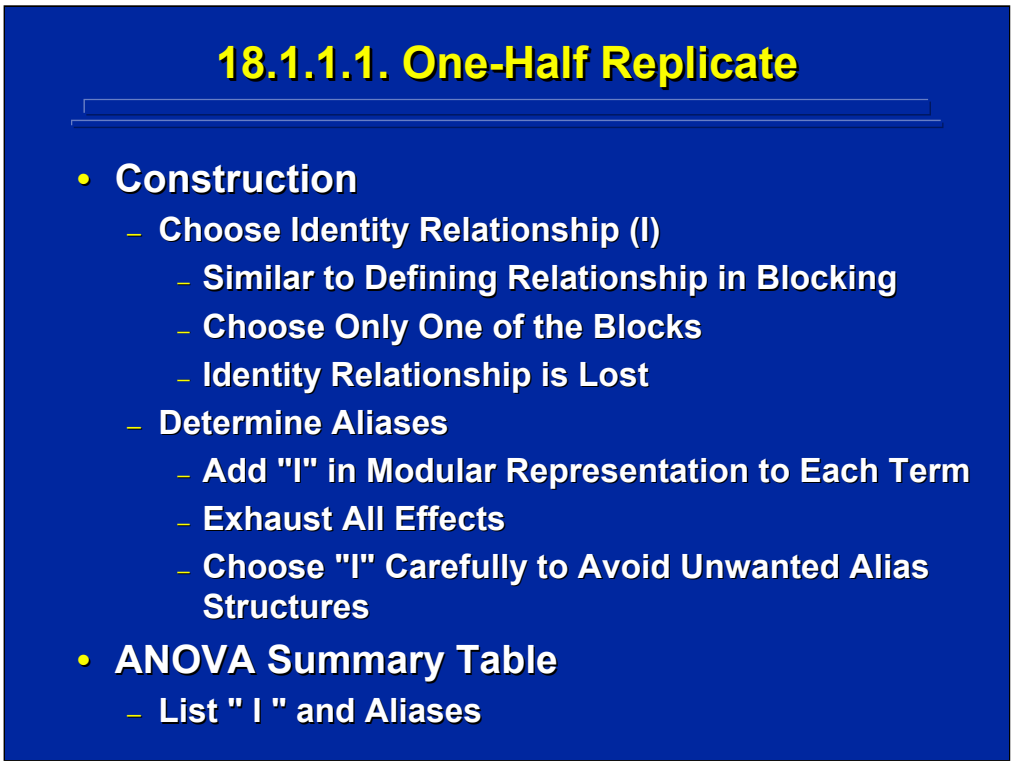



Fractional replicates are primarily used when the full factorial design is too large, and only a subset of the data can be collected. In addition, fractional replicates can be used as a systematic and efficient way to conduct pre-testing. Fractional replicates are also used as components of other advanced experimental design such as central-composite designs described in Topic 22.

Mostly, factors with two levels are used in fractional-factorial designs because complete 1 df effects rather than partial components of effects are used in the identity relationship and the subsequent alias structure. Two of these $2^{k-p}$ fractional replicates are discussed in this topic. A $2^{k-1}$ fractional-factorial design is a one-half replicate of a complete $2^k$ factorial design formed using only one identity relationship. A $2^{k-2}$ fractional-factorial design is a one-quarter replicate of a complete $2^k$ factorial design formed using two identity relationships.

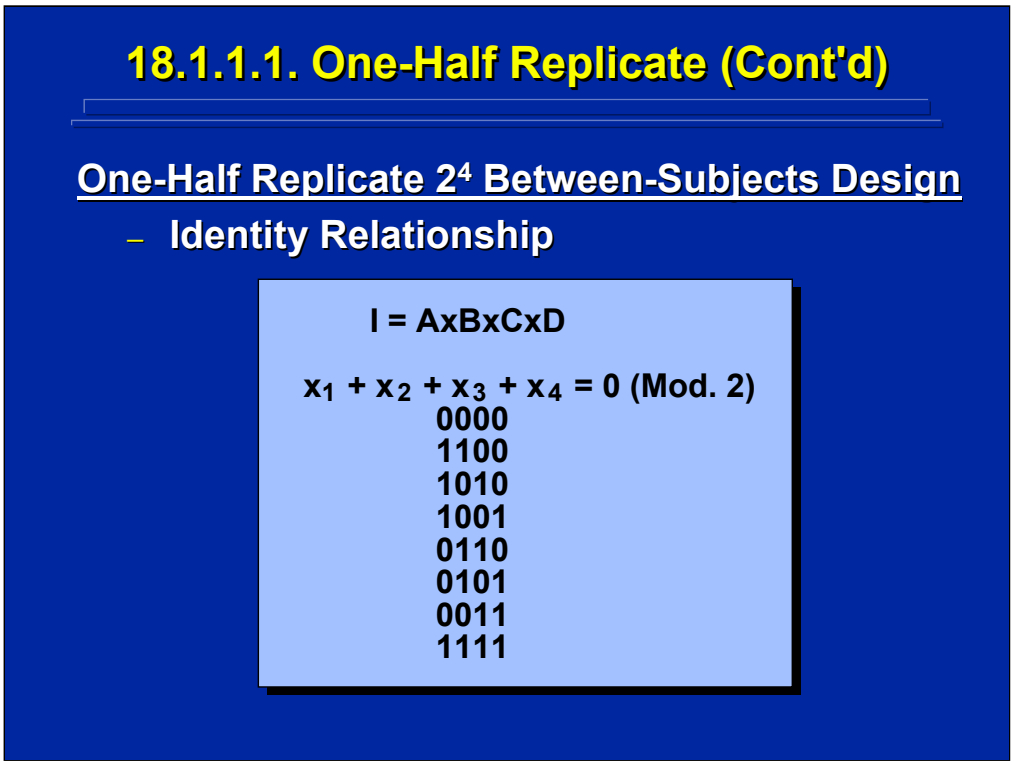



This subsection describes the procedures for construction of one-half and one-fourth replicates of full $2^k$ factorial designs. The procedures for forming one-half replicates are similar to simple blocking, and the procedures for forming one-quarter replicates are similar to complex blocking as described in Topic 17.

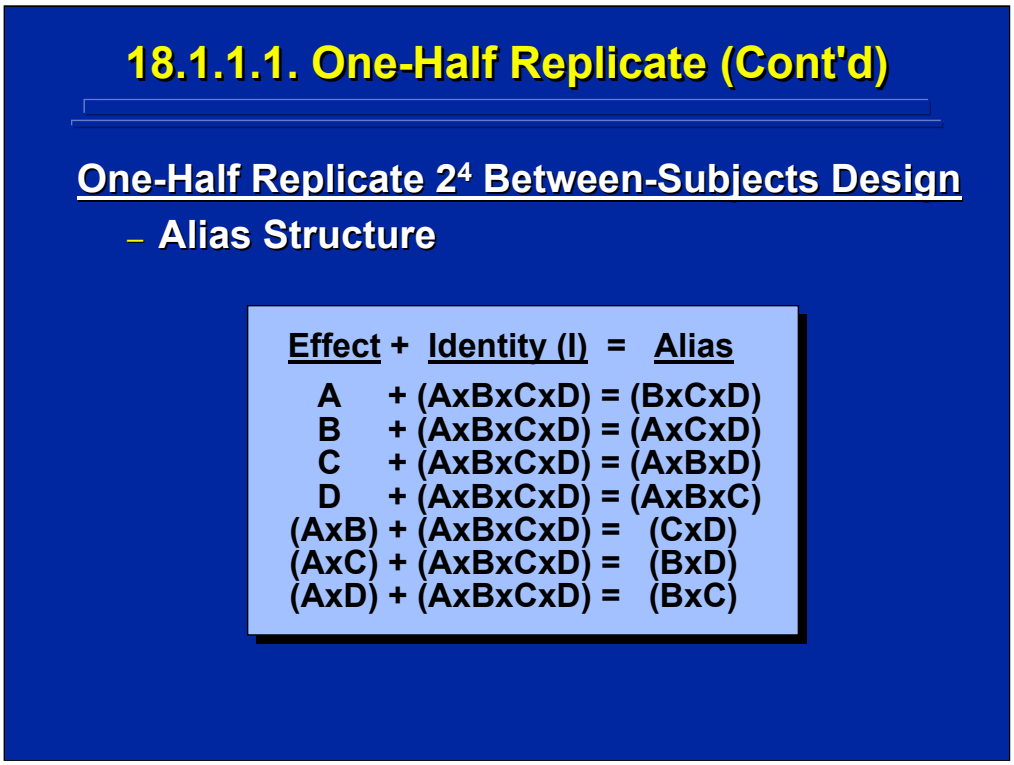



The identity relationship, I, is one level of an effect that is used to determine which half of the treatment conditions in the full $2^k$ factorial design that will be observed in the one-half replicate. This is equivalent to using one level of the defining relationship in Mod. 2 used in simple blocking as a means of defining the one-half replicate of the $2^k$ design. Since the identity relationship is held constant at one level, that effect cannot be tested in the subsequent ANOVA on the one-half replicate. In addition, the choice of the defining relationship also determines which effects in the full factorial design will be confounded with each other in the one-half replicate. One simply adds the defining relationship in Mod. 2 notation to each effect in the full factorial design to determine the confounding effect or alias structure in the one-half replicate. Consequently, the experiment must choose the defining relationship carefully to insure that effects of research interest are not lost or aliased with other effects of interest in the full factorial design.

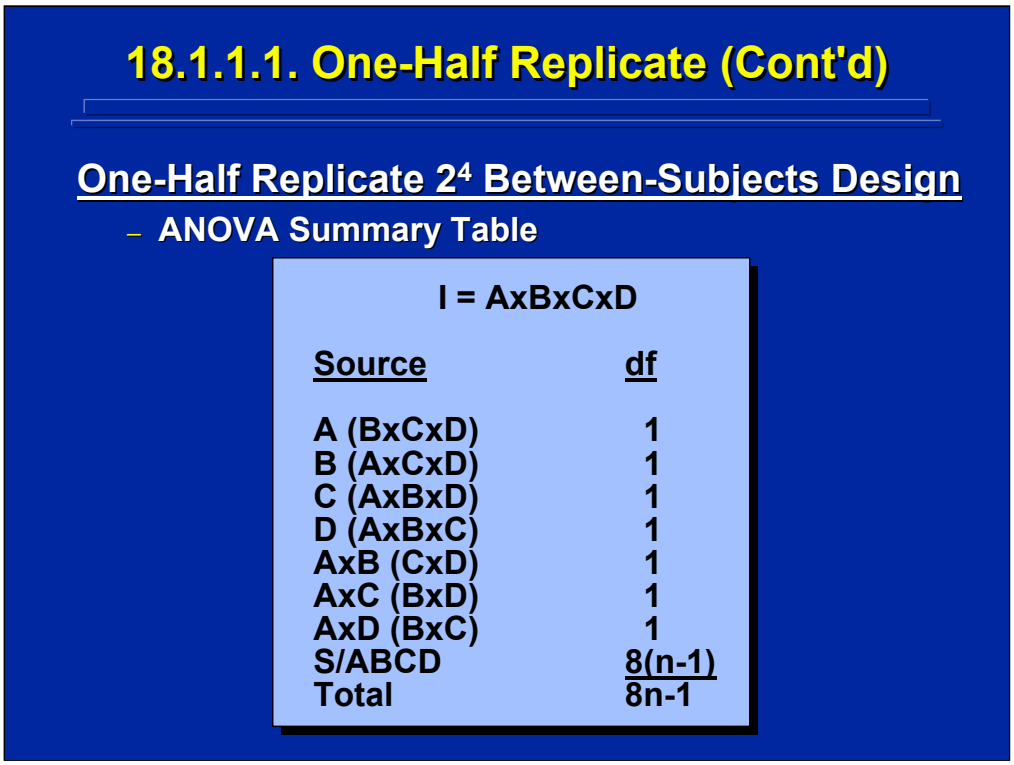
The resulting ANOVA summary table should list both the identity relationship, I, and the alias structure. This allows one to determine how all the effects in the complete $2^k$ factorial design are distributed in the one-half replicate.




This slide shows an example of choosing the AxBxCxD interaction of a $2^4$ factorial design as the defining relationship to split the 16 treatment combinations into a one-half replicate of 8 treatment combinations. Note that the 0 level in Mod. 2 is used as the value of the defining value in choosing the 8 treatment conditions shown on this slide in Mod. 2 notation. Alternatively, the 1 value in Mod. 2 could have been chosen instead to select the other one-half replicate still using AxBxCxD as the defining relationship. Either half-replicate can be used to define the $2^{4-1}$ design where I = AxBxCxD.

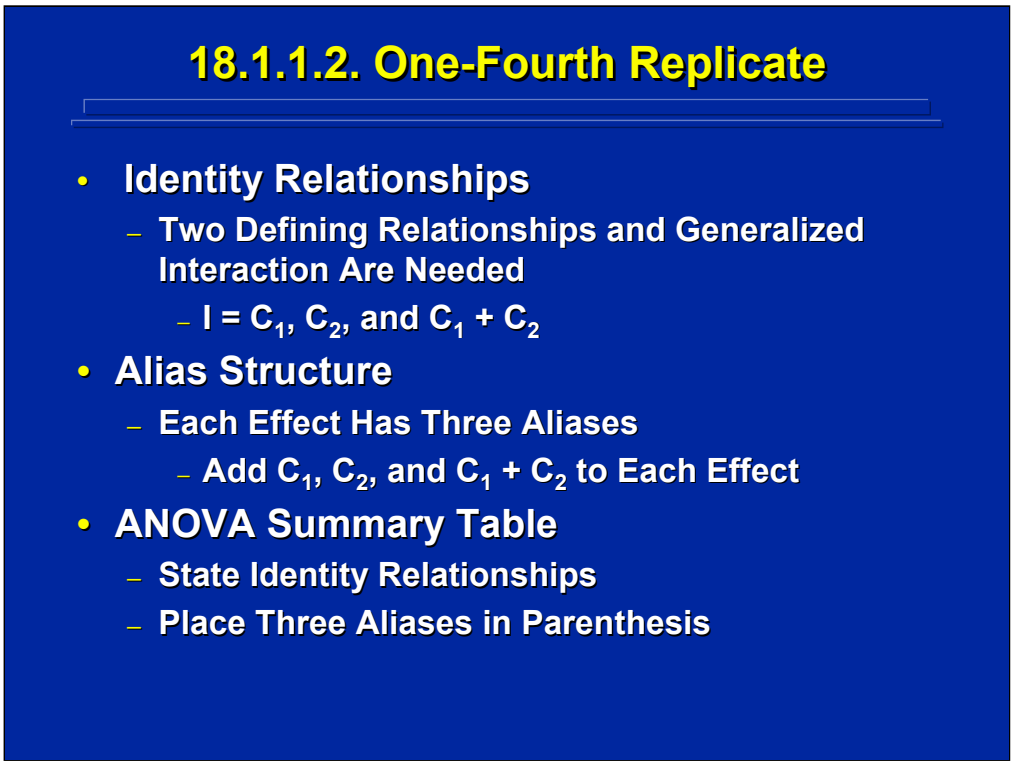



This slide shows the complete alias structure of the one-half replicate of the $2^4$ factorial design when I = AxBxCxD. The identity relationship is added in Mod. 2 to each of the effects in the $2^4$ factorial design until all the effects are exhausted to determine the confounded effects in the one-half replicate. For example, the A main and the BxCxD interaction are totally confounded in this $2^{4-1}$ fractional-factorial design. Notice that each effect in a one-half replicate is confounded with one other effect. The effect assumed not to exist is called the alias.

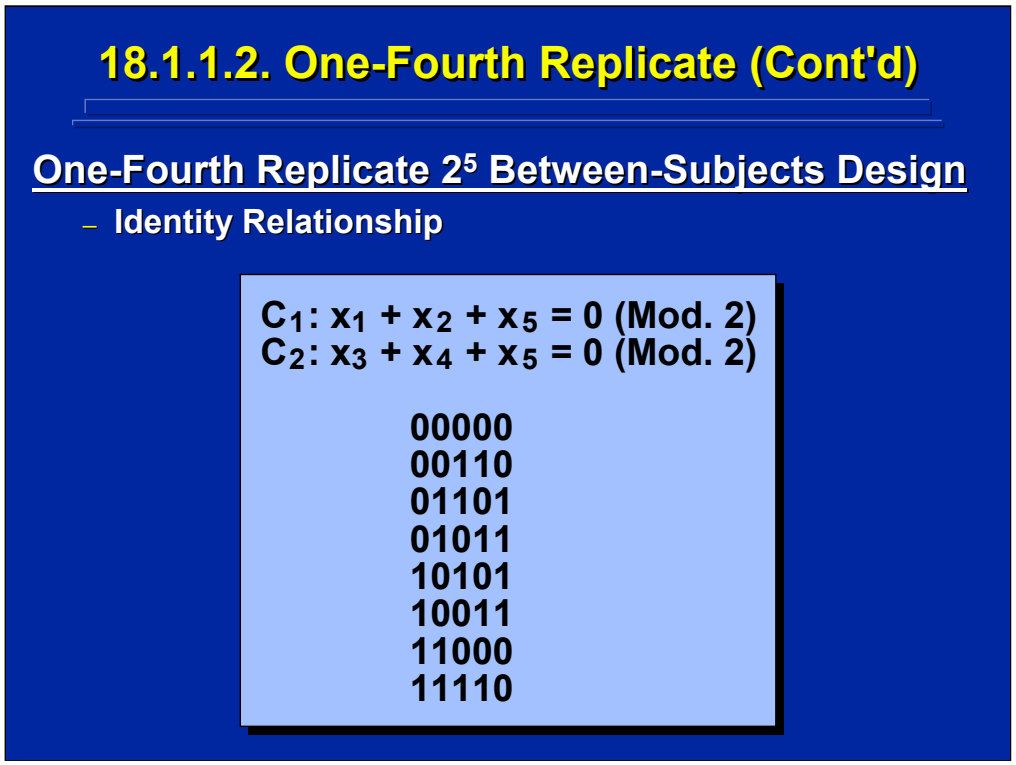



The resulting Sources and degrees of freedom for the example one-half replicate are shown on this slide. Note that the identity relationship is written at the top. The eight treatments in the one-half replicate provides tests of seven 1 df effects. The seven effects and their aliases as determined in the previous slide are listed under Sources. The error term for this between-subjects design is S/ABCD.

Note that the identity relationship, the seven sources, and their aliases account for all 15 effects possible in the full $2^4$ factorial design. There is no test of the AxBxCxD interaction since it is held constant. One must assume that the alias effects listed in parenthesis are negligible and that the resulting F-test represents the effect stated not its alias. For example, if the F-test on the first source listed is significant, the experimenter assumes that Factor A is significant, not the BxCxD interaction alias. In human factors research it is reasonable to assume that main effects rather than the three-way interactions exist. The last three treatment effects in the Source listing show confounding between two-way interactions. It is difficult to determine which should be listed as the effect and which should be the alias unless the experimenter has prior scientific literature information to support the choice. One would have to complete the factorial design to provide tests on each separate effect. At least this one-half replicate can evaluate the four main effects if the full $2^4$ factorial design cannot be conducted.




Two defining relationships are needed to define a one-fourth replicate of a $2^k$ factorial design using Mod. 2 notation. The first identity relationship, $C_1$, divides the factorial design in half, and the second identity relationship, $C_2$, divides each half into two parts to yield the one-fourth replicate. Each effect in the resulting $2^{k-2}$ design has three aliases since the complete identity relationship is defined as I = $C_1$, $C_2$, and $C_1$ plus $C_2$. Consequently, the resulting ANOVA Summary Table would list three aliases for each source.




This slide presents an example of using two defining relationships to split the 32 treatment combinations of the $2^5$ factorial into a one-quarter replicate of 8 treatment combinations. The first relationship, $C_1$, uses the AxBxE interaction and the second relationship, $C_2$, uses the CxDxE interaction. Both $C_1$ and $C_2$ are set at the 0 value in Mod 2. to determine the resulting eight treatment conditions shown in Mod. 2 notation for this one-fourth replicate. This procedure is analogous to using one level of each defining relationship as in complex blocking to specify one of the four sets of eight treatments to be used as the one-fourth replicate of the complete $2^5$ factorial design.

## 18.1.1.2. One-Fourth Replicate (Cont'd)

**One-Fourth Replicate $2^5$ Between-Subjects Design**

- ANOVA Summary Table

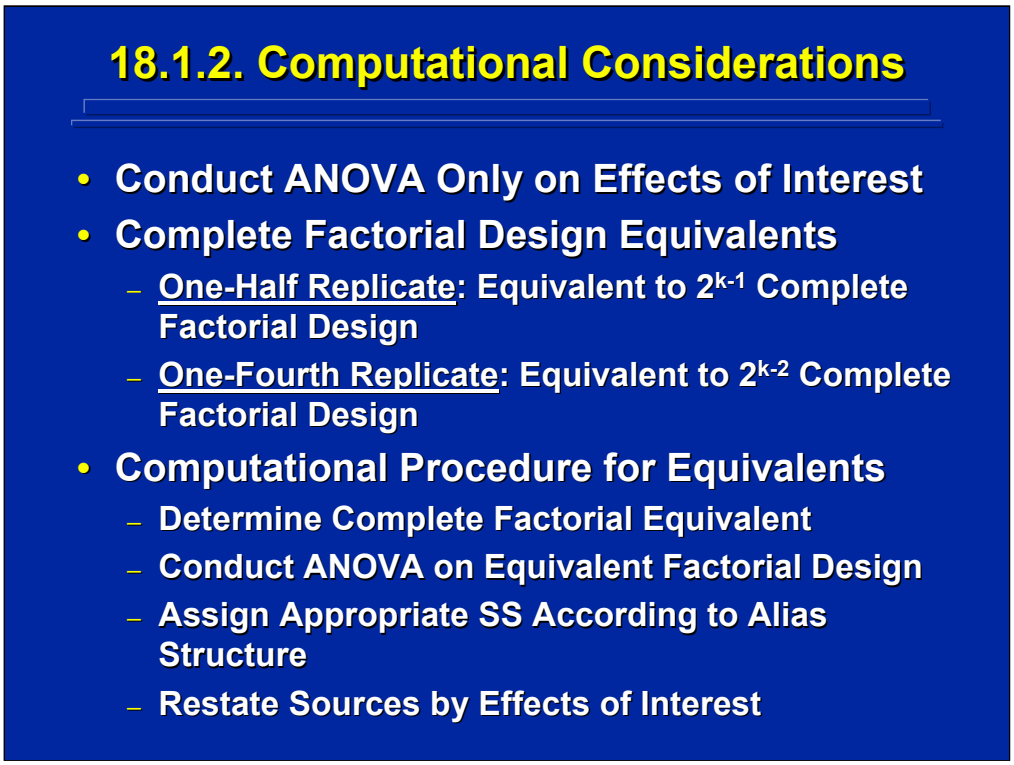
I = AxBxE, CxDxE, AxBxCxD

| Source | df |
|---|---|
| A (BxE, AxCxDxE, BxCxD) | 1 |
| B (AxE, BxCxDxE, AxCxD) | 1 |
| C (AxBxCxE, DxE, AxBxD) | 1 |
| D (AxBxDxE, CxE, AxBxC) | 1 |
| E (AxB, CxD, AxBxCxDxE) | 1 |
| AxC (BxCxE, AxDxE, BxD) | 1 |
| BxC (AxCxE, BxDxE, AxD) | 1 |
| S/ABCDE | 8(n-1) |
| Total | 8n-1 |

The resulting between-subjects Sources and degrees of freedom for the example one-fourth replicate are shown on this slide. Note that the identity relationship is written at the top and includes three interactions, AxBxE, CxDxE, and AxBxCxD, that cannot be evaluated in this design. The eight treatments in the one-fourth replicate provide tests of seven 1 df effects.

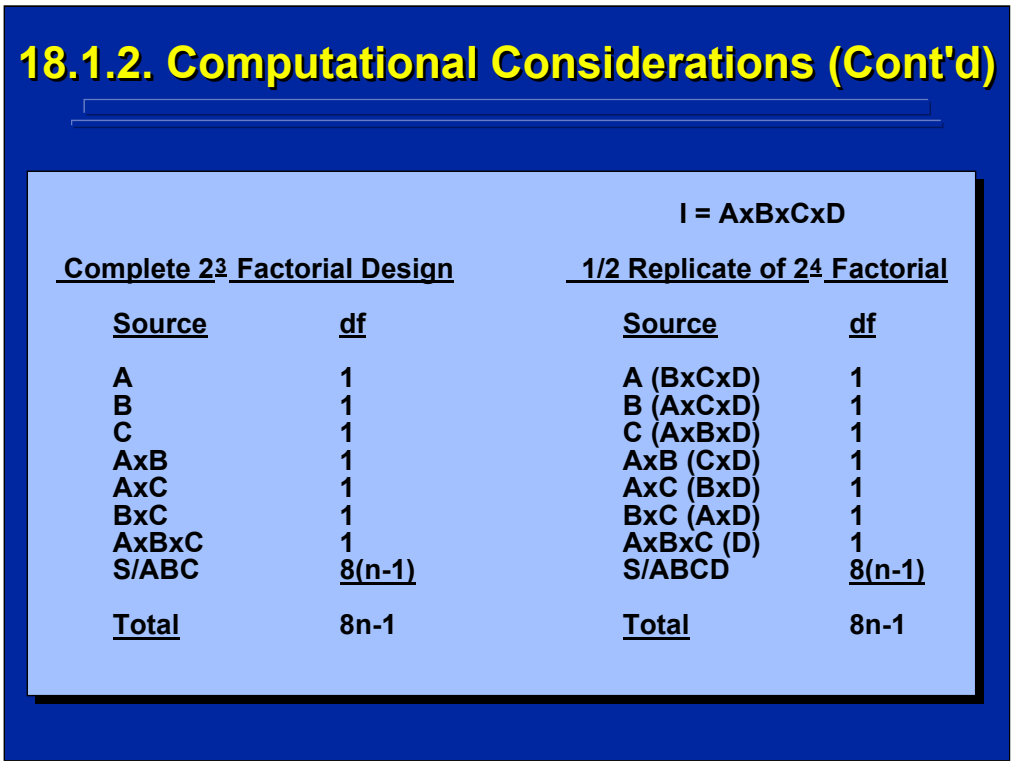
The seven effects and their aliases are listed under Sources. The alias structure is determined by adding each of the three identity interactions to each effect in Mod. 2 notation to exhaust all effects. Again, effects in the identity relationship, sources tested, plus aliases equal all the effects in the complete 25 factorial design. The error term for this between-subjects design is S/ABCDE.

Note that this example one-fourth replicate only allows separate evaluation of each of the five main effects of the full 25 factorial design. Given the small number of treatment observation in the 25-2 design, this is the best resolution of effects from the full factorial design that can be achieved. Again, the experimenter must always determine the identity relationships and alias structure before choosing the fractional replicate to avoid inadvertent confounding of important effects in the experiment.




The most straightforward approach to conducting the ANOVA on fractional-factorial designs is to use standard basic ANOVA rules, procedures, and algorithms to calculate and test the effects of interest. All identity relationships cannot be tested and aliases are ignored due to confounding.

Alternatively, one can consider using the complete factorial design equivalent for calculations. Remember that a one-half replicate of a $2^k$ design is equivalent to a complete $2^{k-1}$ factorial design, and a one-quarter replicate of a $2^k$ design is equivalent to a $2^{k-2}$ factorial design. One could conduct the ANOVA on the equivalent full factorial design, assign the SS according to the alias structure, and restate the Sources as the effects of interest.




This slide demonstrates that a complete $2^3$ factorial design is equivalent to a one-half replicate of a $2^4$ factorial design (i.e., a $2^{4-1}$ or $2^3$ design). The Sources and df of a one-half replicate of a $2^4$ design is shown on the right side of this slide. The four-way interaction is the identity relationship that forms the alias structure and this interaction cannot be estimated in the ANOVA. The Factor D main effect and the two- and three-way interactions including Factor D are listed as aliases. Note that the resulting effects of Factors A, B, and C are exactly the same as the Sources listed on the left side of the slide for the complete $2^3$ factorial design.

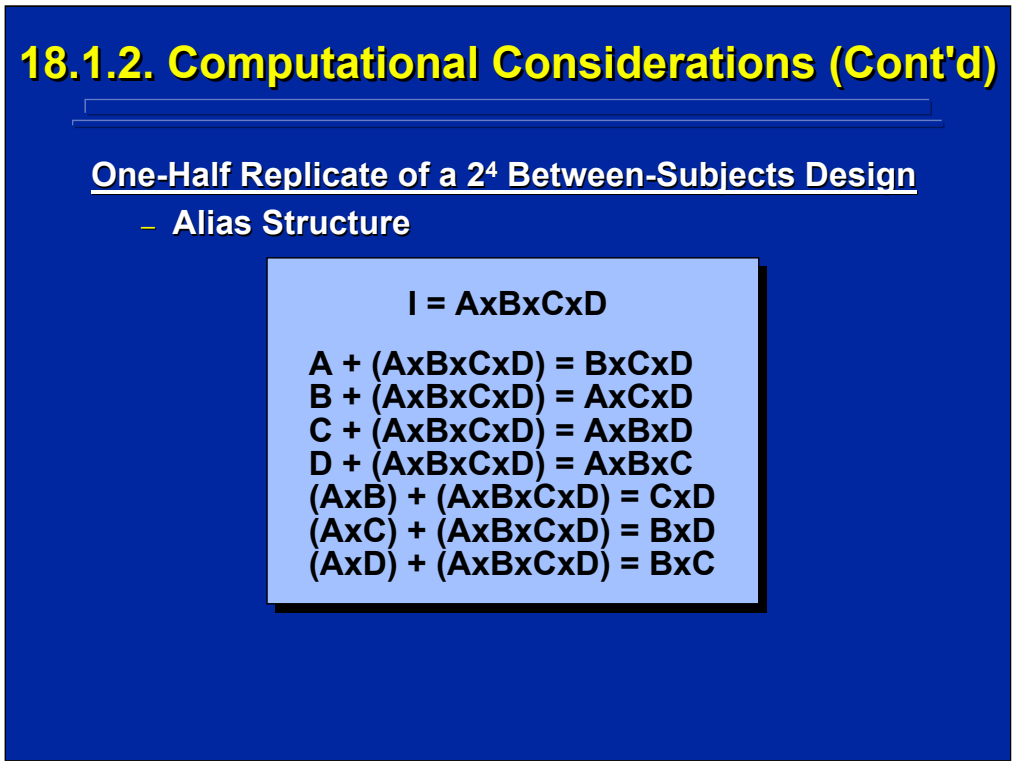
Usually the experimenter would reverse the alias statement AxBxC (D) and list Factor D as the effect of interest by restating it as D (AxBxC) in the one-fourth replicate ANOVA. Nevertheless, calculating either Factor D or the AxBxC interaction would yield the same SS value because these two effects are totally confounded in the one-fourth replicate.

## 18.1.2. Computational Considerations (Cont'd)

- **Example Problem. Preliminary testing was conducted on a new computerized target detection system. Two different settings of four different factors including target speed (A), target size (B), noise level (C), and display resolution (D) were evaluated. Five different soldiers completed 100 detection trials in only one treatment combination of the four factors tested to calculate the percent of targets detected. A one-half replicate of the full factorial design was used to pretest main effects and the existence of possible two-way interactions. Do the settings of any of the four main effects of target factors and two-way interactions have a significant effect on the percent of targets detected? ($p < 0.01$)**

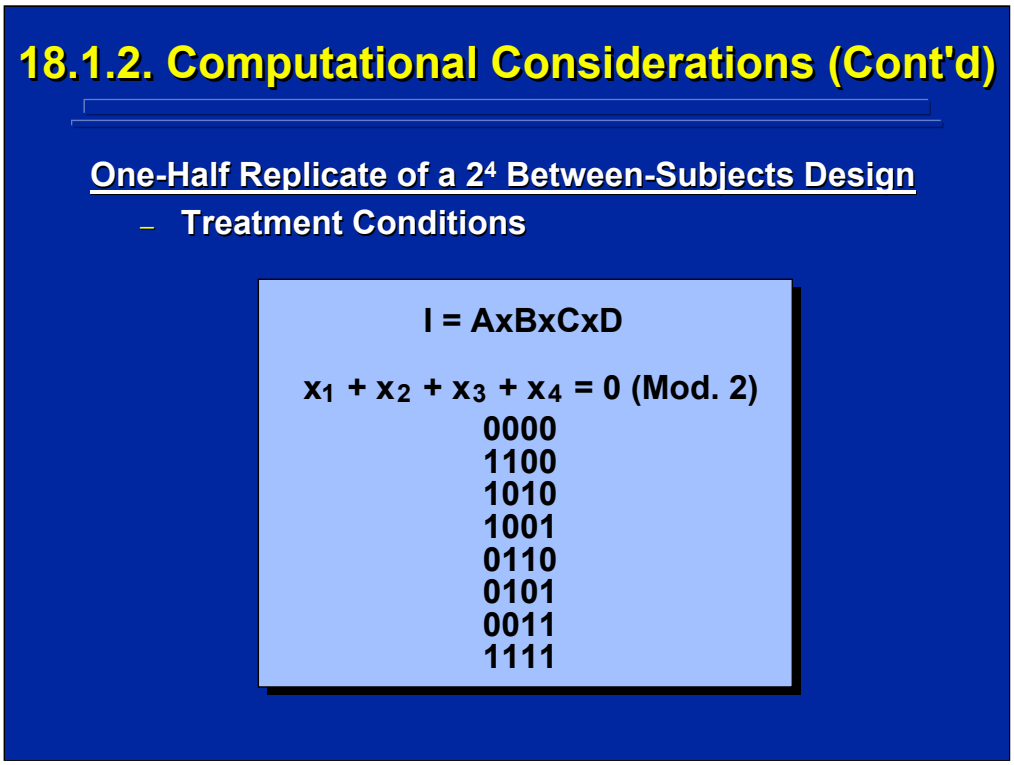
(Click in this red rectangle to see SAS calculations for this example.)

This example describes a one-half replicate of a $2^4$ between-subjects target detection problem where the fractional-factorial design is used to conduct pre-testing. Since this is a between-subjects design, a total of 40 different soldiers are needed for preliminary testing. The Slater and Williges (2006) appendix provides the SAS program for conducting the ANOVA on this fractional-factorial example problem.




(Click in this red rectangle to see SAS calculations for this example.)

This slide lists the complete alias structure of the one-half replicate of the $2^4$ factorial design when the fourth-order interaction, AxBxCxD, is chosen as the identity relationship. Note that all four main effects are not confounded with each other in this design and can be tested separately. Three groups of two-way interactions are confounded. But, the possible existence of two-way interactions can also be evaluated in this preliminary test even though the exact relationship cannot be determined. Third- and fourth-order interactions are purposefully selected for the identity relationship and the alias structure since they are confounded or lost in this design in order to conduct pretest on possible main effects and two-way interactions.




(Click in this red rectangle to see SAS calculations for this example.)

The resulting eight treatment conditions of the one-half replicate are listed on this slide in Mod. 2 notation. The 0 level of the AxBxCxD interaction in Mod. 2 notation is used as the defining relationship to generate the eight treatment combinations.

## 18.1.2. Computational Considerations (Cont'd)

**One-Half Replicate of a $2^4$ Between-Subjects Design**

- Percent of Targets Detected

| $A_0$ | | | | $A_1$ | | | |
|---|---|---|---|---|---|---|---|
| $B_0$ | | $B_1$ | | $B_0$ | | $B_1$ | |
| $C_0$ | $C_1$ | $C_0$ | $C_1$ | $C_0$ | $C_1$ | $C_0$ | $C_1$ |
| $D_0$ | $D_1$ | $D_1$ | $D_0$ | $D_1$ | $D_0$ | $D_0$ | $D_1$ |
| 0000 | 0011 | 0101 | 0110 | 1001 | 1010 | 1100 | 1111 |
| 0.50 | 0.78 | 0.70 | 0.02 | 0.28 | 0.75 | 0.09 | 0.14 |
| 0.23 | 0.89 | 0.67 | 0.43 | 0.22 | 0.48 | 0.23 | 0.27 |
| 0.45 | 0.64 | 0.90 | 0.14 | 0.39 | 0.38 | 0.14 | 0.08 |
| 0.66 | 0.50 | 0.87 | 0.27 | 0.08 | 0.89 | 0.37 | 0.31 |
| 0.37 | 0.40 | 0.76 | 0.19 | 0.44 | 0.66 | 0.46 | 0.25 |

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the hypothetical data of the $2^{4-1}$ fractional-factorial design. The two levels of each of the four factors, A, B, C, and D, are listed in the top four rows of this slide. The resulting eight treatment combinations shown on the previous slide are underlined and listed in Mod. 2 notation in a middle row on this slide. The percent of targets detected for each of the 40 different soldiers participating in this pretest are listed in the bottom five rows of this slide representing five soldiers tested in each of the eight treatment combinations of the between-subjects, one-half replicate design.

## 18.1.2. Computational Considerations (Cont'd)

**One-Half Replicate of a $2^4$ Between-Subjects Design**

- **ANOVA Summary Table**

I = AxBxCxD

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Speed (A) [BxCxD] | 1 | 0.2993 | 0.2993 | 12.32 * |
| Size (B) [AxCxD] | 1 | 0.1823 | 0.1823 | 7.50 * |
| Noise (C) [AxBxD] | 1 | 0.0029 | 0.0029 | 0.12 |
| Resolution (D) [AxBxC] | 1 | 0.0865 | 0.0865 | 3.56 |
| AxB [CxD] | 1 | 0.0774 | 0.0774 | 3.19 |
| AxC [BxD] | 1 | 0.2822 | 0.2822 | 11.61 * |
| BxC [AxD] | 1 | 0.8526 | 0.8526 | 35.09 * |
| S/ABCD | 32 | 0.7776 | 0.0243 | |
| Total | 39 | 2.5608 | | |

*p < 0.01

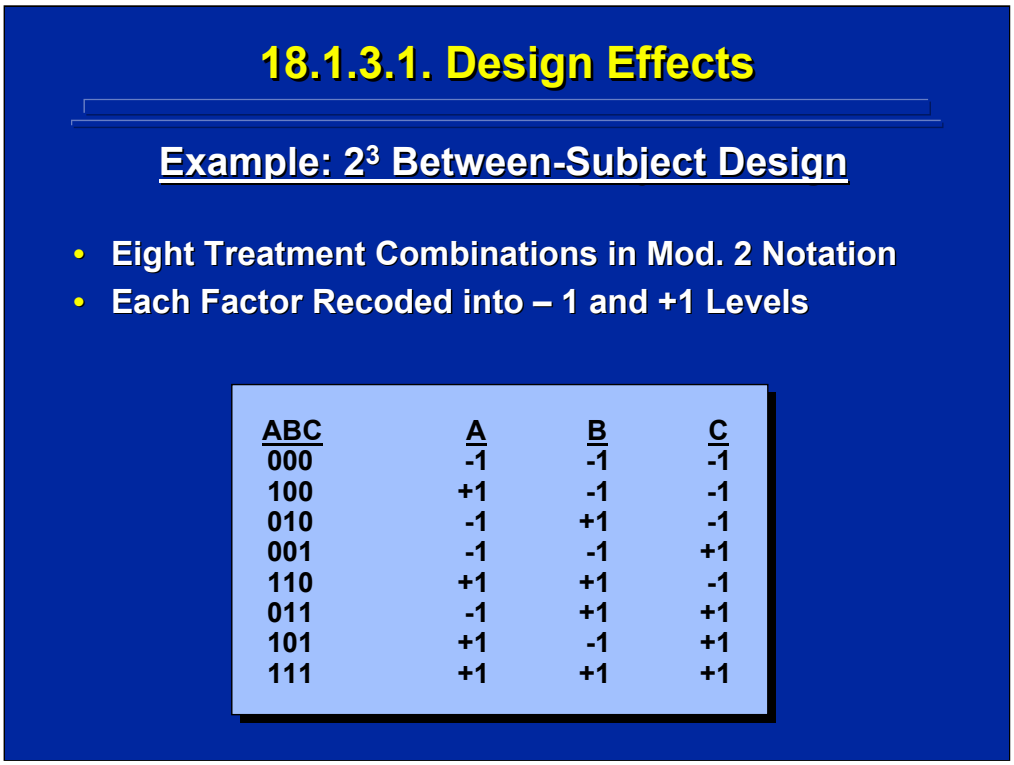
(Click in this red rectangle to see SAS calculations for this example.)

The complete ANOVA summary Table for the example $2^{4-1}$ fractional-factorial design is shown on this slide. The identity relationship is listed at the top of the slide, and aliases are listed in brackets besides each of the seven effects that are tested. Since this is a between-subjects design, each effect is tested by S/ABCD assuming Subjects are random and the four factors are fixed-effects factors.

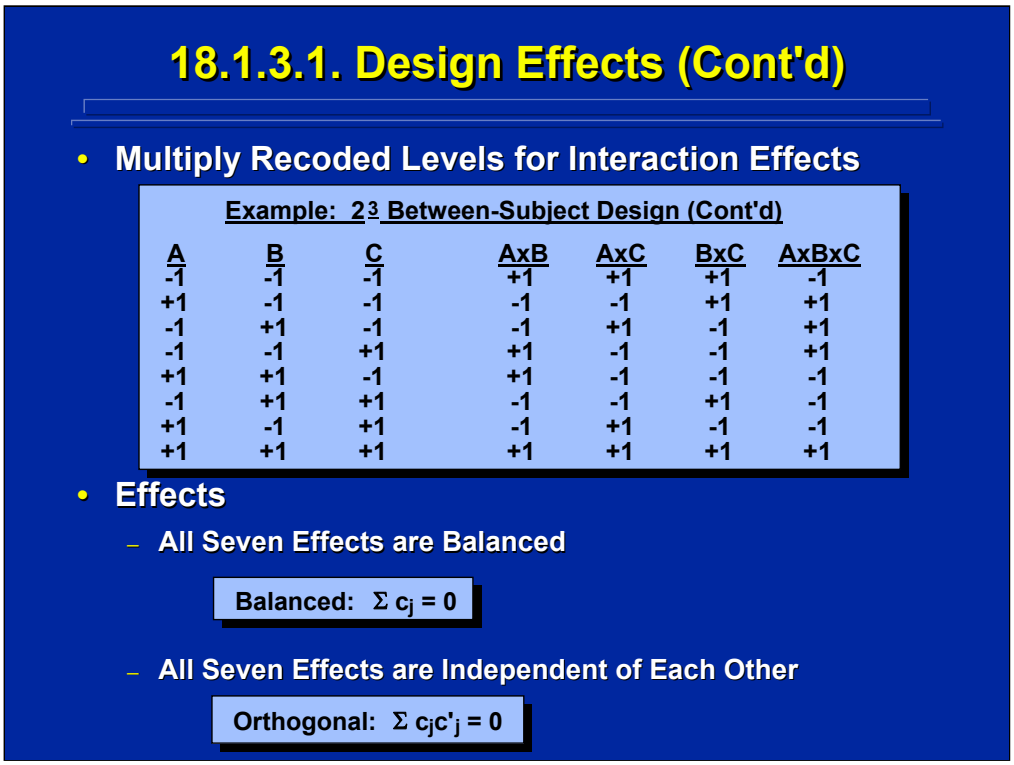
Standard ANOVA procedures are used to calculate the ANOVA from the data set shown on the previous slide. A complete $2^3$ factorial design could be conducted, and the values for the AxBxC interaction are restated as the main effect Resolution (D) due to the alias structure. Alternatively, just the seven effects, A, B, C, D, AxB, AxC, and BxC could be calculated separately.

The F-tests conducted in this pretest show that the main effects of Target Speed and Target Size significantly ($p < 0.01$) affect the percent of targets detected assuming three-way interactions do not exist. In addition, two of the two-factor interaction groupings, AxC [BxD] and BxC [AxD], are significant. In order to resolve the two-way interactions, the other half of the $2^4$ factorial design must be conducted.

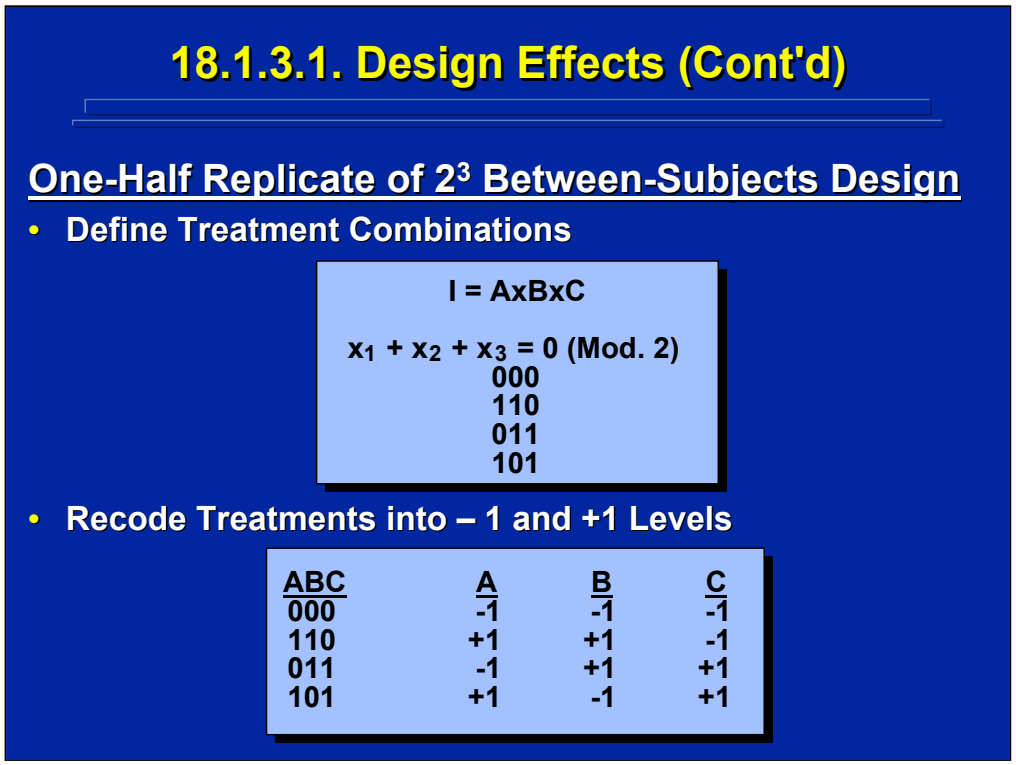
This subsection discusses the concept of design resolution that is present in fractional-factorial designs, and describes how design resolution can be used in choosing appropriate $2^{k-p}$ fractional replicates.



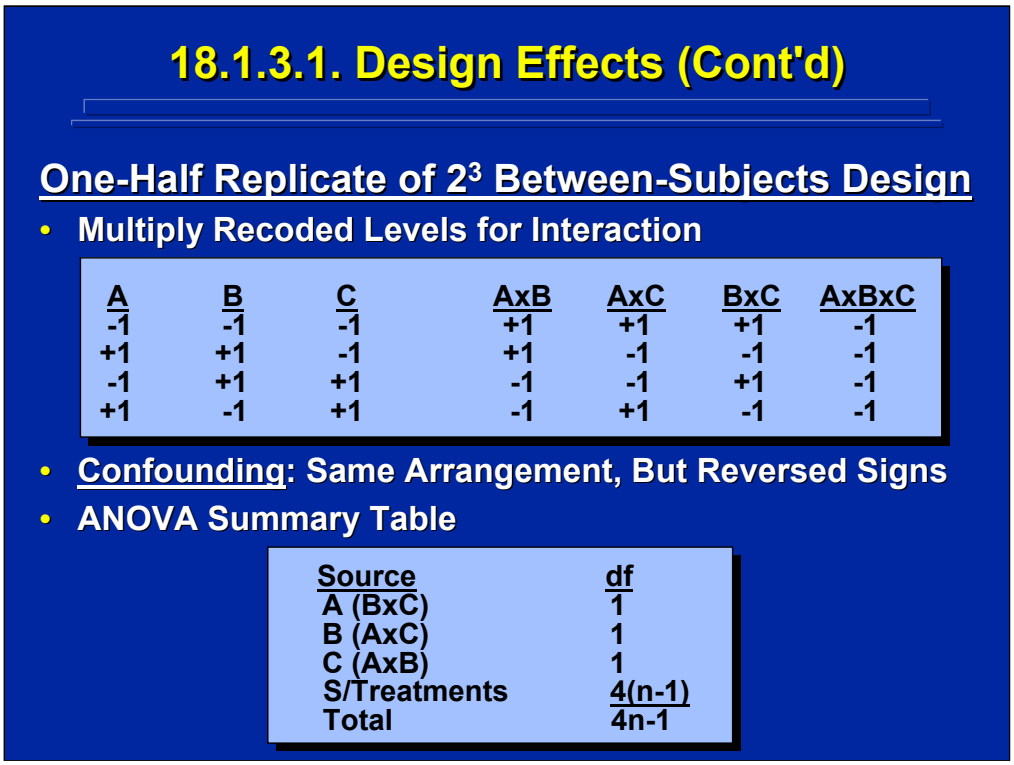

Fractional-factorial designs can be described in terms of all the effects that are present in factorial designs. In a $2^3$ factorial design, there are a total of eight treatment combinations. These eight treatments are listed in Mod. 2 notation in the left column on this slide. Alternatively, the 0 and 1 levels in Mod. 2 can be recoded as -1 and +1, respectively. This recoding for each factor is shown under the A, B, and C column designations on this slide.




The +/-1 levels of each factor are multiplied together to obtain the +/-1 levels of interaction effects in the factorial design as shown on the top part of this slide for the $2^3$ factorial design. All seven effects in this factorial design are both balanced and independent of each other in terms of the +/-1 factor level weighting, c, according to the standard requirements as listed in the formulae at the bottom of this slide.




The top portion of this slide lists the four treatment conditions, in Mod. 2 notation, in a one-half replicate of the $2^3$ factorial design when the 0 level of the three-way interaction is used as the defining relationship. These four treatment conditions are recoded into +/-1 levels of Factors A, B, and C in the lower portion of the slide.




The top of this slide shows the seven effects of the $2^3$ factorial design as represented in the four treatment conditions of the one-half replication shown on the previous slide. Again, the interaction weightings are determined by multiplying the +/-1 weightings of the factors involved.

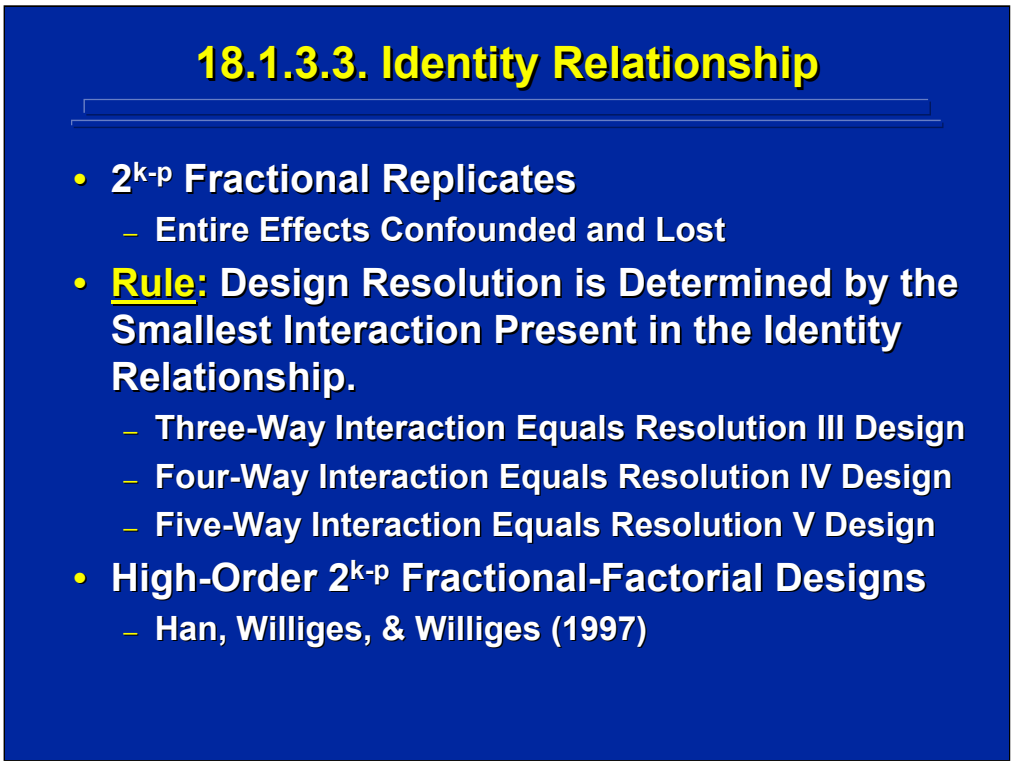
Effects that are confounded have the same column arrangement of + and - signs, but the +/-1 weightings are reversed. For example, the main effect of Factor A (with weighting arrangement -1, +1, -1, and +1) is confounded with the BxC interaction (with weighting arrangement +1, -1, +1, and -1). Notice the AxBxC interaction is held constant at the -1 level in all four treatments because it is the identity relationship. The Sources and degrees of freedom of the resulting ANOVA Summary Table shown at the bottom of this slide show the aliases of A, B, and C that reflect these confounded effects. Consequently, the +/-1 representations clearly show the confounding effects present in $2^{k-p}$ fractional replicates.

## 18.1.3.2. Definitions of Resolution

- **Design Resolution: The quality of information that can be obtained from an experiment as determined by the effects confounded in the alias structure.**
  - **Resolution III: All main effects are unconfounded with each other and can be evaluated assuming all interactions do not exist.**
  - **Resolution IV: All main effects and groups of two-way interactions are unconfounded with each other and can be evaluated assuming all three-way and higher interactions are zero.**
  - **Resolution V: All main effects and two-way interactions are unconfounded with each other and can be evaluated assuming all three-way and higher interactions are zero.**

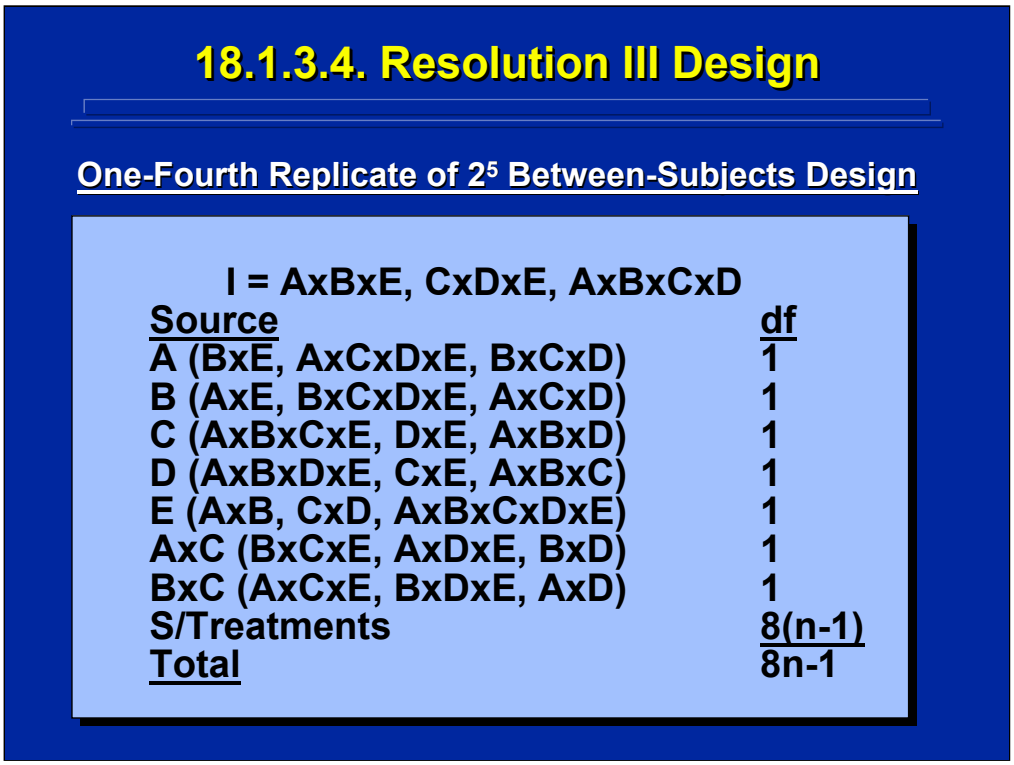
Design resolution is the quality of information that can be obtained from a fractional replicate. All main effects are not confounded with each other in Resolution III designs. All main effects and groups of two-way interactions are unconfounded in Resolution IV designs. And, all main effects and two-way interactions are not confounded with each other in Resolution V designs. So, as resolution increases the quality of unconfounded information increases in fractional replicates.

Obviously, the experimenter is interested in the highest resolution possible in an experimental design. In most human factors and ergonomics research, main effects and two-way interaction are of primary importance which requires a Resolution V design. At times, this is not possible. For example, the highest resolution possible in the one-half replicate shown in the previous slide is a Resolution III design due to the restricted number of resulting treatment conditions in the fractional-factorial design. Nonetheless, design resolution should always guide the experimenter in choosing a fractional replicate alternative.

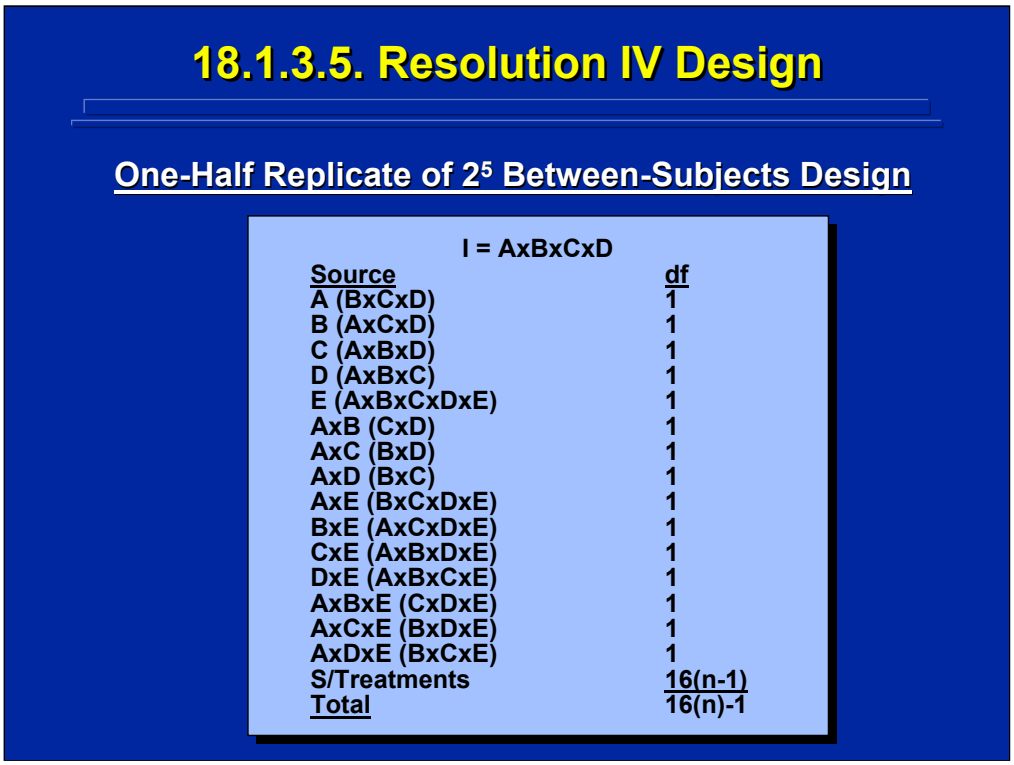



Entire effects are confounded and lost in fractional replicates of $2^k$ designs because each effect has only one degree of freedom. Resolution of $2^{k-p}$ fractional replicates are determined by the smallest interaction in the identity relationship. Therefore, the resolution number equals the smallest interaction present in the identity relationship. For example, a Resolution V fractional replicate has a five-way interaction as the lowest-order effect in the identity relationship. Hence, one can only construct a Resolution V fractional replicate with a minimum of five factors in the $2^k$ factorial.

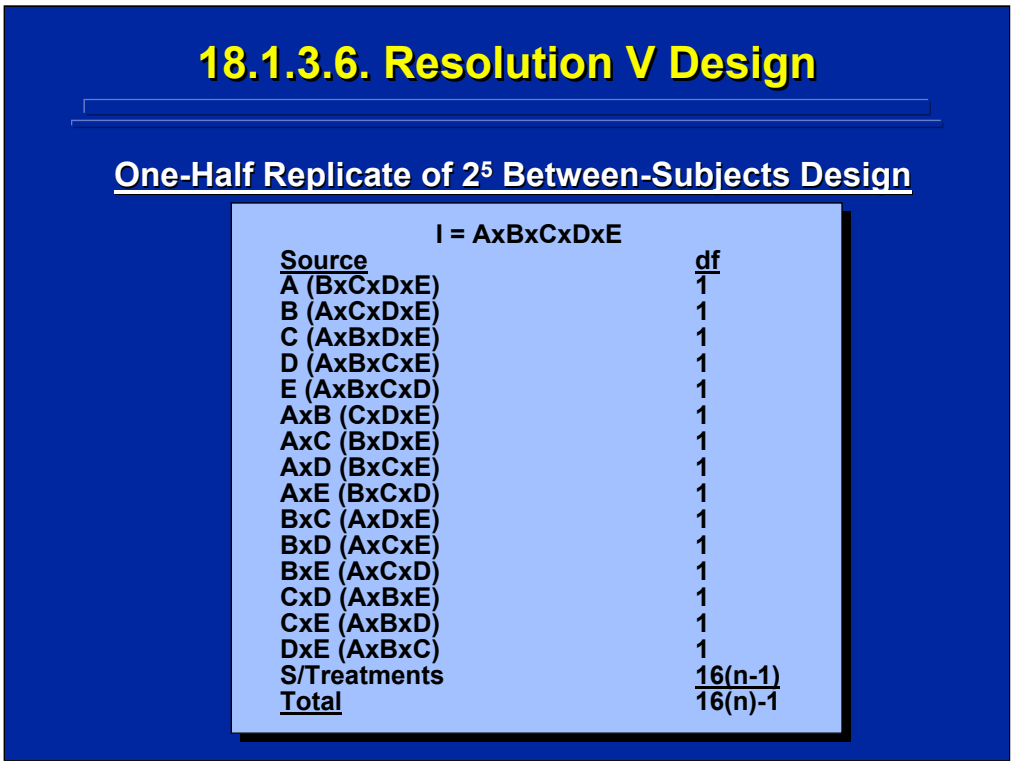
Often, however, more than five factors need to be considered simultaneously in human factors research. Han, Williges, and Williges (1997, page 746) provide the defining relationships of several Resolution III, IV, and V alternatives for $2^{k-p}$ designs in Table 2 that can be used to conduct screening experiments on 5 to 20 factors simultaneously that require only 8, 16, or 32 different treatment conditions.




This slide shows a $2^{5-2}$ fractional replicate. The smallest interaction in the identity relationship is a three-way interaction resulting in a Resolution III design that keeps all five main effects unconfounded as shown in the Source listing. This is the highest resolution possible in a one-fourth replicate of a $2^5$ factorial design.




This slide shows a $2^{5-1}$ fractional replicate. The smallest interaction in the identity relationship is a four-way interaction resulting in a Resolution IV one-half replicate that keeps all five main effects and groups of two-way interactions unconfounded from each other as shown in the Source listing.




This slide shows another example of a $2^{5-1}$ fractional replicate. The smallest interaction in the identity relationship is a five-way interaction resulting in a Resolution V one-half replicate that keeps all five main effects and two-way interactions unconfounded from each other as shown in the Source listing. Obviously, this is a better one-half replicate than the alternative shown on the previous slide because it results in higher design resolution. Consequently, the experimenter must consider the identity relationship carefully before choosing a fractional replicate.

## 18.1.3.7. Uses of Design Resolution

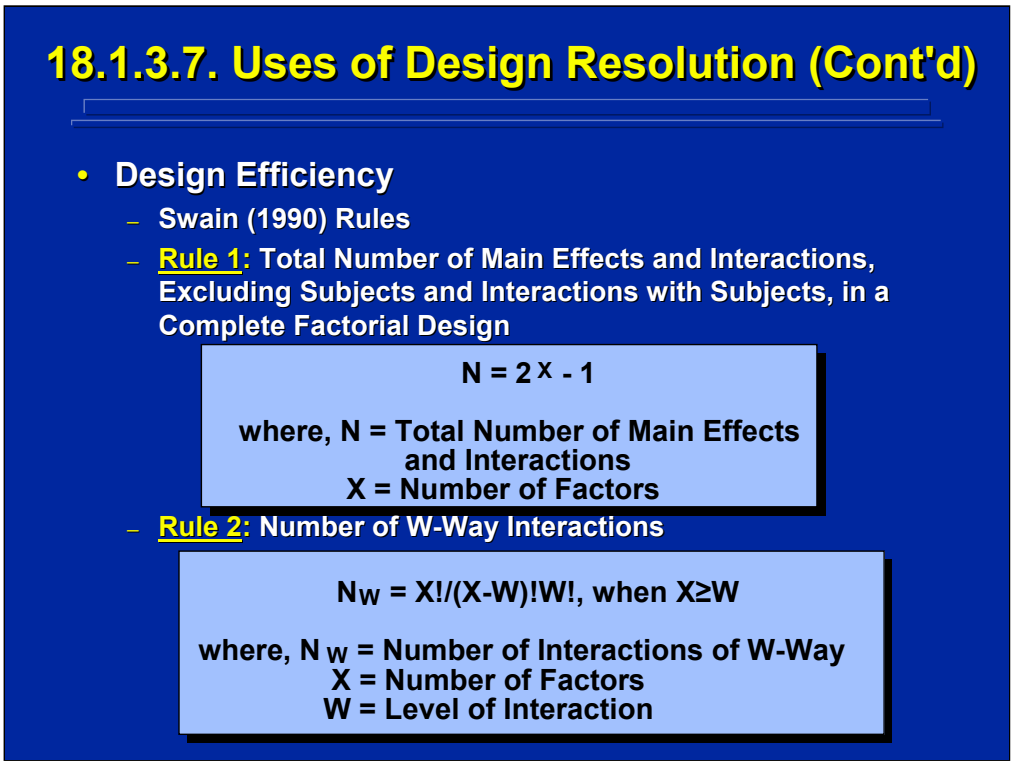
- Set Level of Confounding
  - Interpretation of Results
  - Screening Experiments
- Choice of Fractional-Factorial

Example: One-Half Replicate of $2^4$ Between-Subjects Design

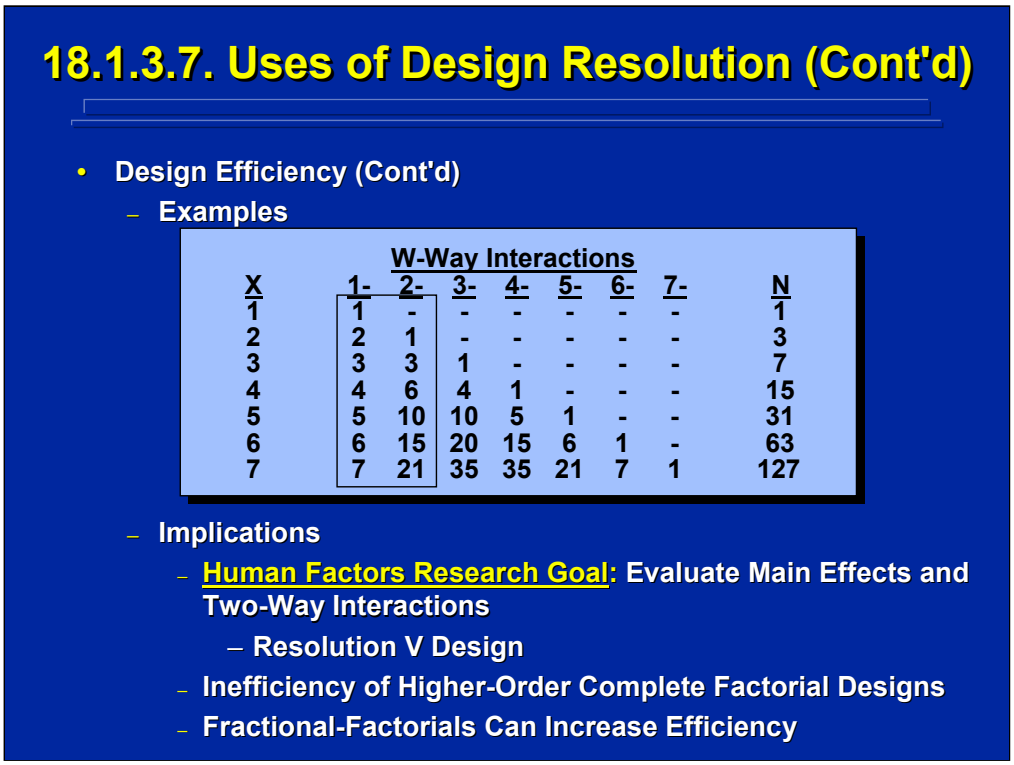
| I = AxBxC | | I = AxBxCxD | |
|---|---|---|---|
| Source | df | Source | df |
| A (BxC) | 1 | A (BxCxD) | 1 |
| B (AxC) | 1 | B (AxCxD) | 1 |
| C (AxB) | 1 | C (AxBxD) | 1 |
| D (AxBxCxD) | 1 | D (AxBxC) | 1 |
| AxD (BxCxD) | 1 | AxB (CxD) | 1 |
| BxD (AxCxD) | 1 | AxC (BxD) | 1 |
| CxD (AxBxD) | 1 | AxD (BxC) | 1 |
| S/Treatments | 8(n-1) | S/Treatments | 8(n-1) |
| Total | 8n-1 | Total | 8n-1 |

Design resolution can be used in several ways in $2^{k-p}$ fractional-factorial designs. First, it can be used to set the level of confounding present in fractional replicates to facilitate interpretation of results and choice of designs for screening experiments. If, for example, the experimenter is primarily interested in main effects, then Resolution III designs can be used. If, on the other hand, the experimenter is interested in evaluating main effects and two-way interactions, a Resolution V design is needed.

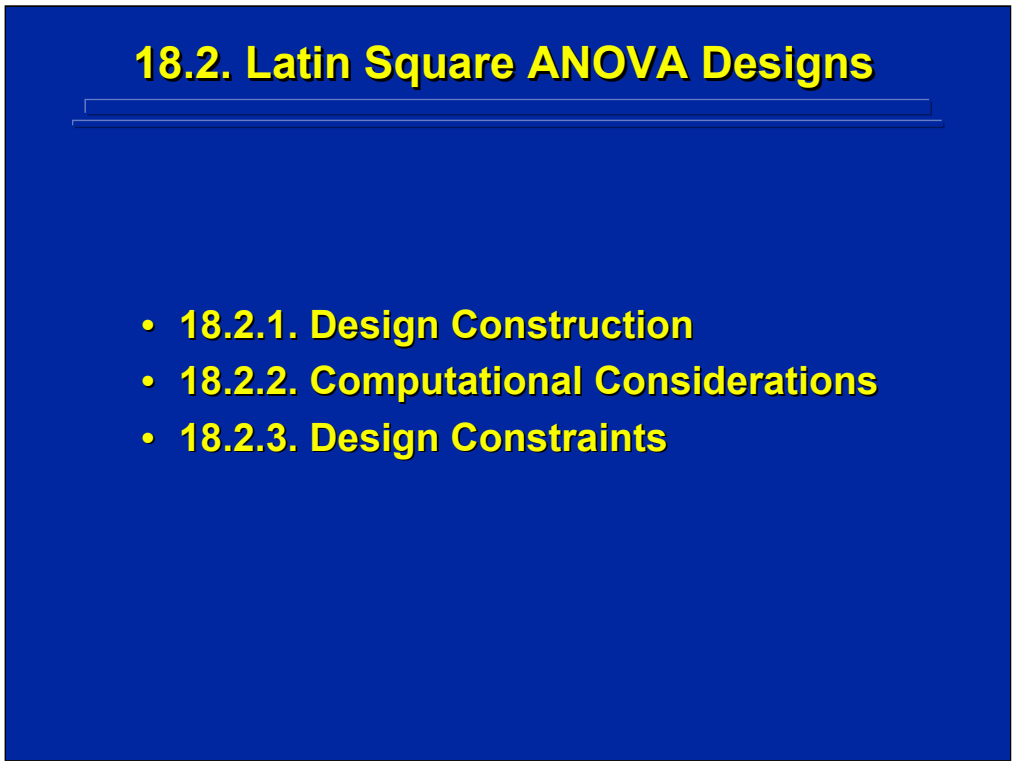
Second, design resolution can be used to assist the experimenter in choosing the defining relationship for any fractional factorial design. The bottom of this slide compares two versions of a one-half replicate of a $2^4$ factorial design. The left side is one possible Resolution III alternative that uses the AxBxC interaction as the identity relationship and the right side is a Resolution IV alternative using the fourth-order interaction as the identity relationship. The Resolution IV alternative is better because none of the four main effects includes two-way interactions as aliases. Consequently, the experimenter should always choose the highest resolution when selecting a fractional-factorial design alternative.




Design resolution can also be used when assessing the design efficiency of large factorial designs. If the experimenter is interested in only main effects and two-way interactions (i.e., Resolution V effects), a $2^k$ complete factorial design becomes inefficient in evaluating these effects as the number of factors, k, increases since data are collected to evaluate many third- and higher-order interactions in the complete factorial design. A fractional-factorial design may be a more efficient alternative in terms of data collection requirements for these higher-order factorial designs. This slide provides two rules developed by Swain (1990) that specify the number of main effects and interactions as well as the number of any particular level interaction in a factorial design.




The table on this slide uses the two Swain (1990) rules presented on the previous slide to calculate the number of main effects, interactions, and total number of effects present in complete factorial designs having one to seven factors. The numbers within the box show the number of main effects and two-way interactions present in these complete factorial designs. Note that once the number of factors is five or greater, the higher-order interactions constitute the majority of the effects evaluated. For example, only 28 of the 127 effects evaluated in a $2^7$ factorial design are main effects and two-way interactions. If the human factors researcher is only interested in main effects and two-way interactions, a Resolution V fractional replicate may be a more efficient design alternative than a complete $2^7$ factorial design in terms of data collection requirements and the number of effects of interest evaluated.




The last subsection of this topic describes a special case of fractional-factorial designs called Latin square designs that can be used to evaluate the main effects of three factors of interest when each factor has the same number of levels. The construction, computational considerations, and constraints of Latin square designs are discussed separately.

## 18.2. Latin Square ANOVA Designs (Cont'd)

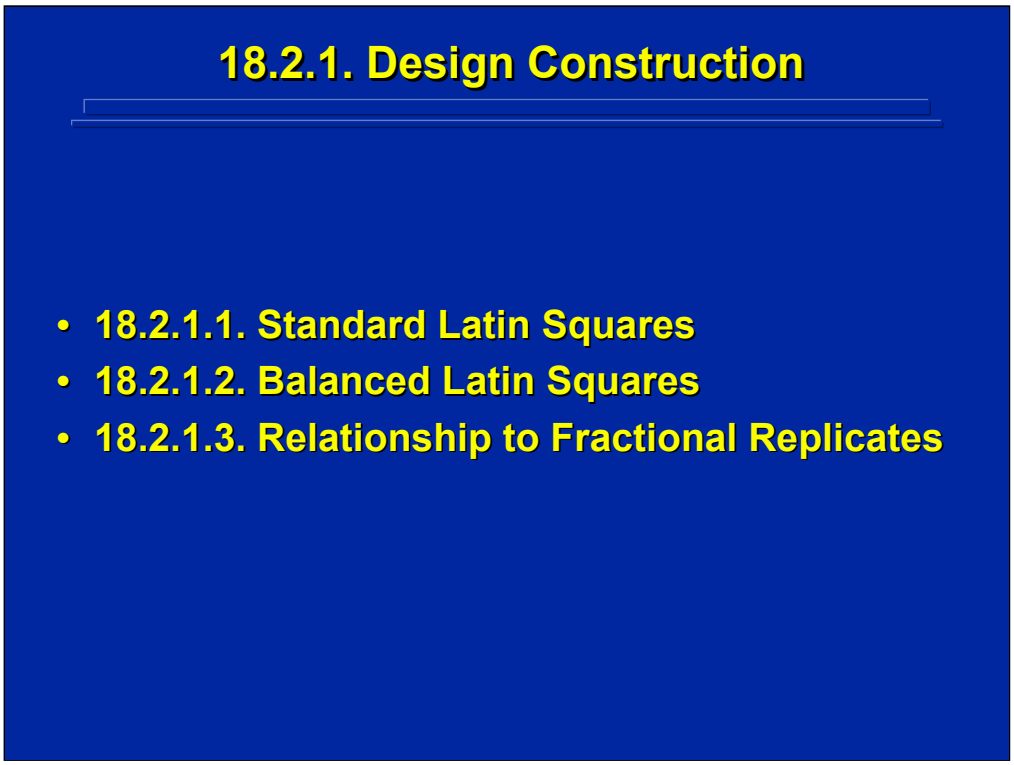
- **Definition**: Three-factor design in which the levels of A appear once in each row (B) and each column (C).
- **Example**: Three Level Latin Square

| | $C_1$ | $C_2$ | $C_3$ |
|---|---|---|---|
| $B_1$ | $A_1$ | $A_2$ | $A_3$ |
| $B_2$ | $A_2$ | $A_3$ | $A_1$ |
| $B_3$ | $A_3$ | $A_1$ | $A_2$ |

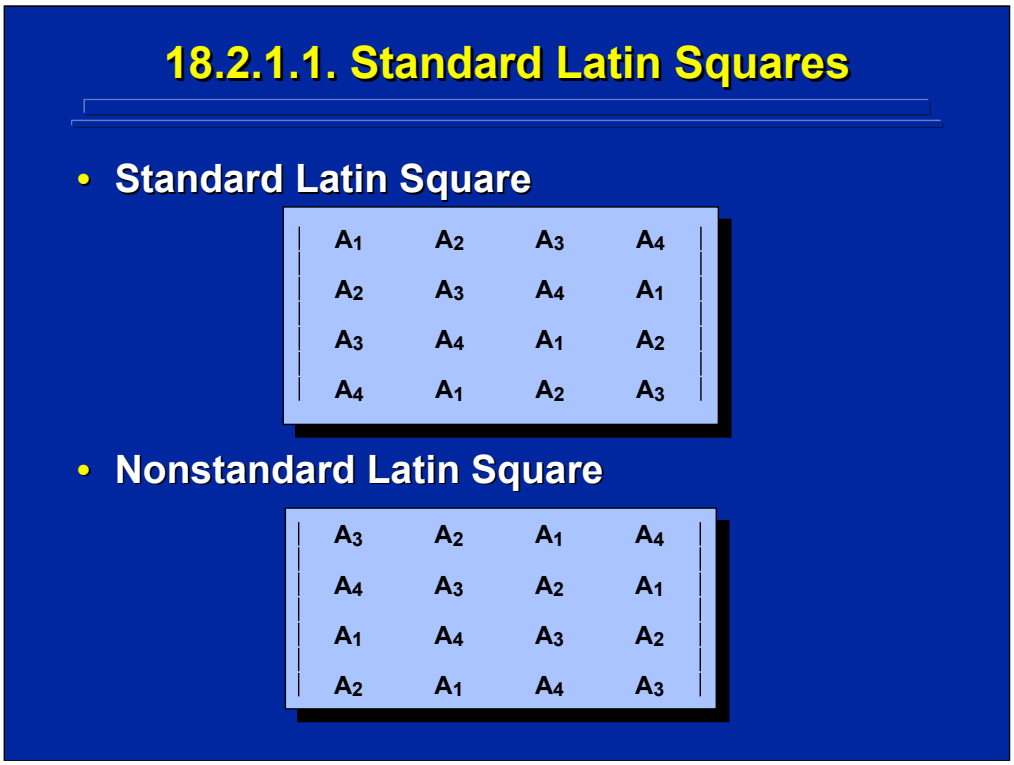
  - Incomplete Factorial Design (9 of 27 Treatments)
  - Several Latin Squares possible

Every Latin square design is a three-factor design in which the levels of factor A appear once in each row (Factor B) and once in each column (Factor C). All three factors in the Latin square design have the same number of levels.

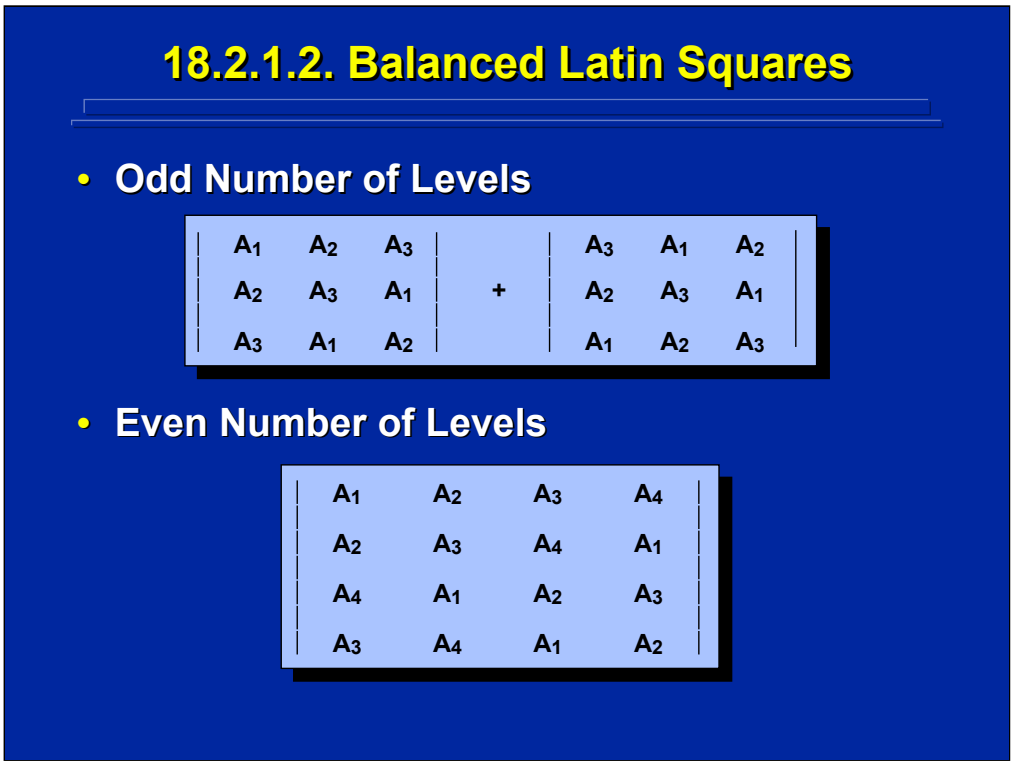
For example, a three-level Latin square is shown in the center of this slide. Each of the nine treatments is defined by the specific combination of levels subscripted for Factors A, B, and C, respectively. Note that these nine treatments only represent one-third of the 27 treatment combinations in the complete $3^3$ factorial design. This example is just one of several 3x3 Latin square designs that are possible. Each Latin square provides enough data to evaluate only the main effects of the three factors, not any of the interactions among them.




Although many Latin squares are possible, two types of Latin squares, standard and balanced, have special characteristics. In addition, Latin square designs can be viewed as a special case of fractional-factorial designs.

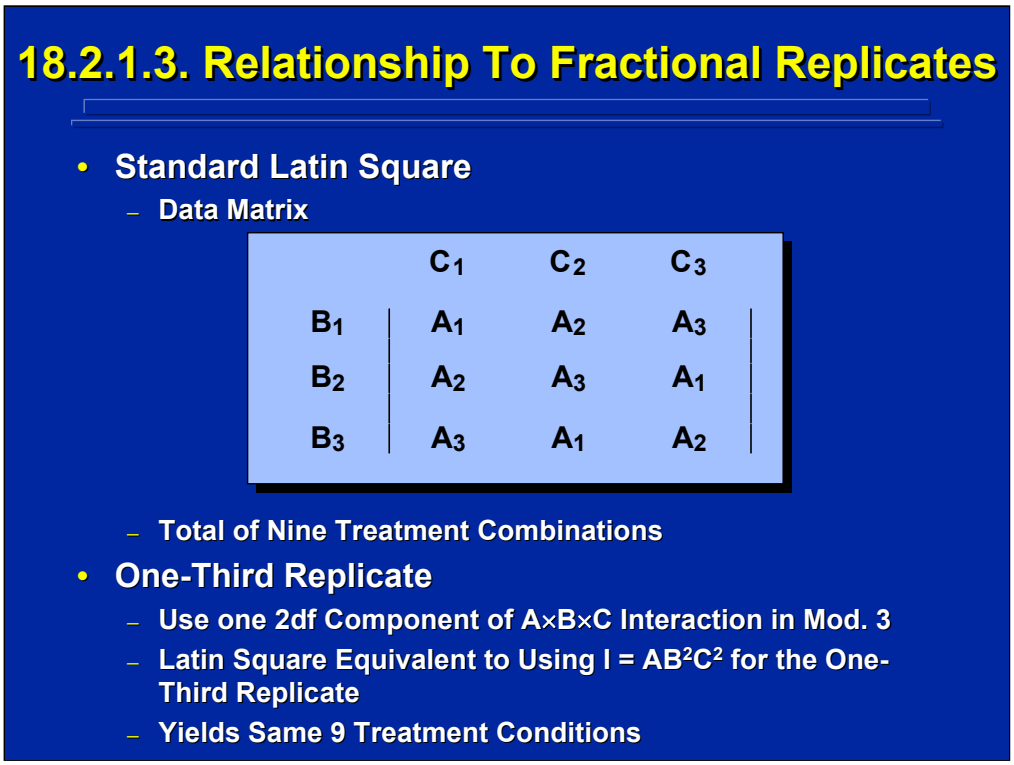



A 4x4 Latin square shown on the upper part of this slide is called “standard” because the first row and column of Factor A levels are in numerical order. Note that each level of Factor A appears in each row (Factor B levels) and column (Factor C levels) only once as required in any Latin square design. The nonstandard version of the 4x4 Latin square shown in the lower part of this slide is one in which the first row and column of Factor A levels are not in numerical order.

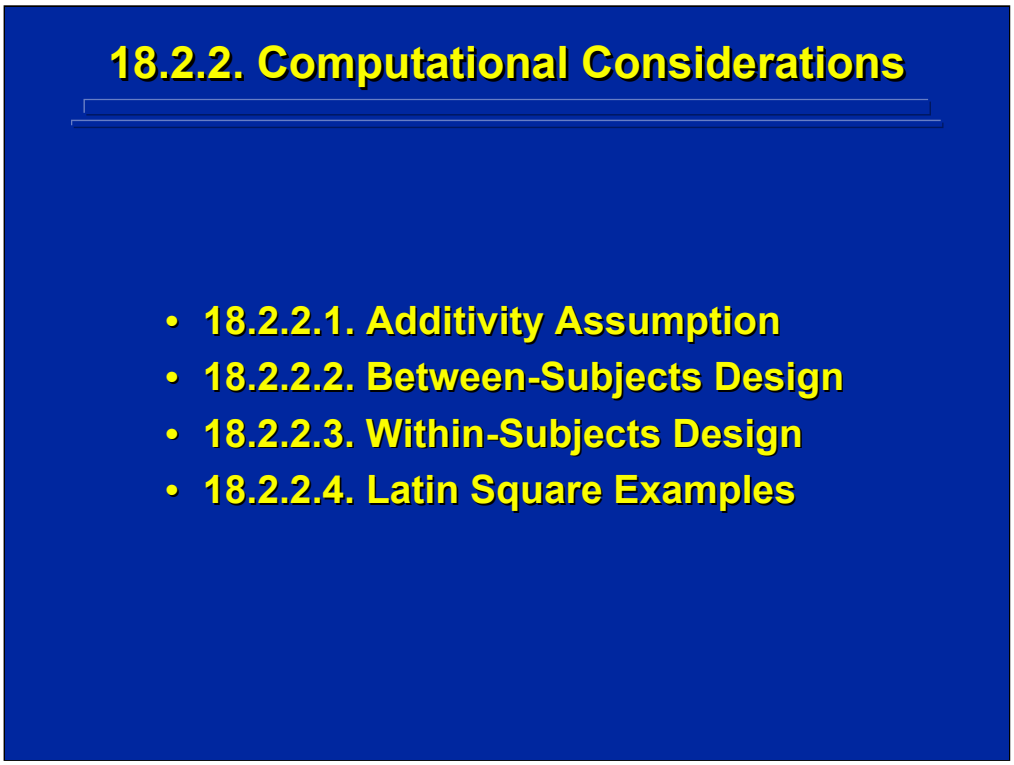



The balanced Latin square is a special case of nonstandard Latin square designs in which each level of Factor A precedes and follows the other levels of Factor A an equal number of times. Rules for generating and analyzing balanced Latin squares are presented in Topic 12 as a means of partially counterbalancing treatment conditions (Factor A) across subjects (Factor B) and presentation order (Factor C) in within-subjects designs.

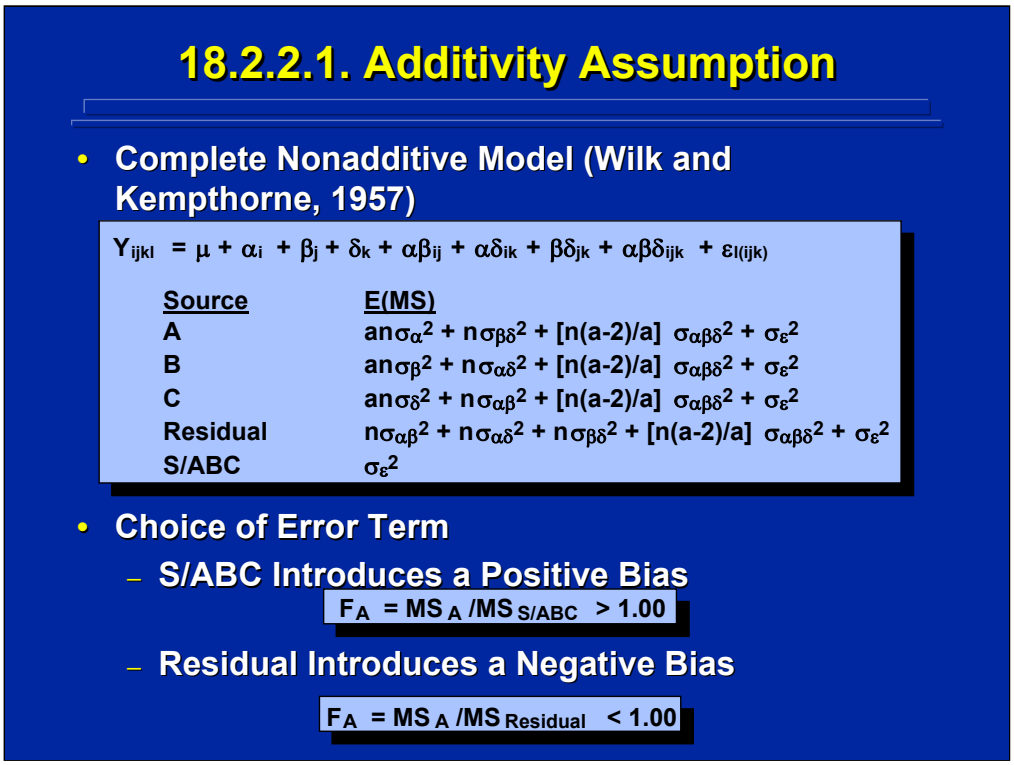
If the number of levels of Factor A are odd, two Latin squares are needed for balancing as shown in the 3x3 Latin squares on the top of this slide. The first 3x3 Latin square is standard and the second 3x3 Latin square is nonstandard that begins with the inverse of the first column and row. Every level will follow and precede every other level twice. If the number of levels of Factor A are even, only one nonstandard Latin square is needed such that every level of Factor A follows and precedes every level once across the design. A 4x4 balanced Latin square design is shown at the bottom of this slide.




Latin square designs are special cases of fractional-factorial designs. The 3x3 standard Latin square design shown on this slide has nine treatment conditions and is equivalent to a one-third replicate of a $3^3$ factorial design. A two degree of freedom component of the AxBxC interaction can be used to form a one-third replicate If the $AB^2C^2$ component, in Mod. 3 notation, of the AxBxC interaction is used to generate a one-third replicate, the resulting nine treatment conditions would be the same as the nine treatment conditions of the 3x3 standard Latin square design shown on this slide.




Due to the reduced number of treatment conditions and the resulting confounding of effects, only the main effects of Factors A, B, and C can be tested in Latin square designs. An additivity assumption that assumes the presence of no interactions in the statistical model is needed to construct F-ratios for these designs. Both between-subjects and within-subjects versions of Latin square computations are described along with a computational example using both versions of a Latin square design.




The expected mean squares, E(MS), for the complete nonadditive statistical model of a Latin square design were derived by Wilk and Kempthorne (1957) and summarized in the top portion of this slide. Note that the nonadditive model includes the interactions among Factors A, B, and C, in the E(MS) due to the confounding present in Latin square designs. The Residual source of variance is the remaining composite variance after the three main effects are calculated. The resulting E(MS) values for the three factors, A, B, C, and the Residual are listed for this design.

Given these E(MS) values, the choice of the appropriate error term for testing each of the three main effects is problematic. The standard between-subjects design error term, S/ABC, introduces a positive bias in the F-test; whereas, the Residual error term introduces a negative bias in the F-test.

## 18.2.2.1. Additivity Assumption (Cont'd)

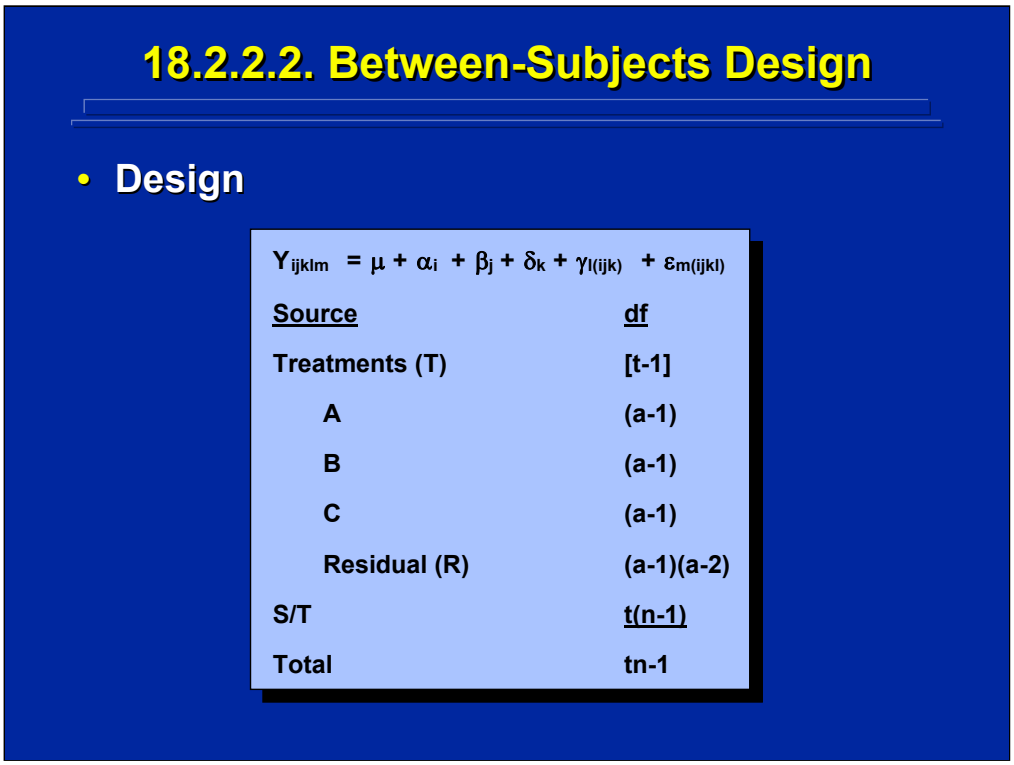
- Additivity Assumption: Interactions Do Not Exist in Model

| Source | E(MS) |
|---|---|
| A | $an\sigma_\alpha^2 + \sigma_\varepsilon^2$ |
| B | $an\sigma_\beta^2 + \sigma_\varepsilon^2$ |
| C | $an\sigma_\delta^2 + \sigma_\varepsilon^2$ |
| Residual | $\sigma_\varepsilon^2$ |
| S/ABC | $\sigma_\varepsilon^2$ |

  - Residual and S/ABC are Appropriate Error Terms
  - Use Residual if No Replication Exists
  - Use Standard Error Term, S/ABC, if available
  - Pooled Error Term of S/ABC and Residual Combined
    - Test Residual by S/ABC for Additivity Assumption
    - Pool if Not Significant ($p < 0.20$)

The additivity assumption means no interactions exist among Factors A, B, and C. If no interactions are present in the statistical model, the E(MS) for the Latin square design presented on the previous slide reduce to the values listed on this slide, and either Residual or S/ABC would be an appropriate error term for testing each of the three main effects.

The Residual effect is used as the error term when replication does not exist. When a balanced Latin square is used for counterbalance within-subjects designs as described in Topic 12, the Residual is used as the error term to test the order main effect. The usual procedure, however, is to use the standard error term, S/ABC, when replication is present. Alternatively, the experimenter could use a pooling procedure by first testing Residual by S/ABC. This preliminary test is a test of the additivity assumption since only interaction effects can be present in the Residual component. If Residual is not significant, when tested at a high $\alpha$ level to guard against Type II error, then S/ABC and Residual can be combined into a pooled error term.




The statistical model under the additivity assumption for a between-subjects Latin square design is stated at the top of this slide. This slide also summarizes the sources and degrees of freedom for a between-subjects Latin square design. The Factor A, B, and C main effects and the Residual effect are merely listed under treatments, because the sum of squares of these sources add to the total treatment effect (T). The effect S/T is the standard S/ABC effect in a between-subjects design. Since all three factors in a Latin square design have the same number of levels, the degrees of freedom for each factor are simply listed as a-1 and the degrees of freedom of Residual can be listed as (a-1)(a-2).

## 18.2.2.2. Between-Subjects Design (Cont'd)

- **Sum of Squares Computational Formulae**

$SS_T = (\sum ABC_{ijk}.^2/n) - (T....^2/tn)$

$SS_A = (\sum A_{i}...^2/an) - (T....^2/tn)$

$SS_B = (\sum B._{j}..^2/an) - (T....^2/tn)$

$SS_C = (\sum C.._{k}.^2/an) - (T....^2/tn)$

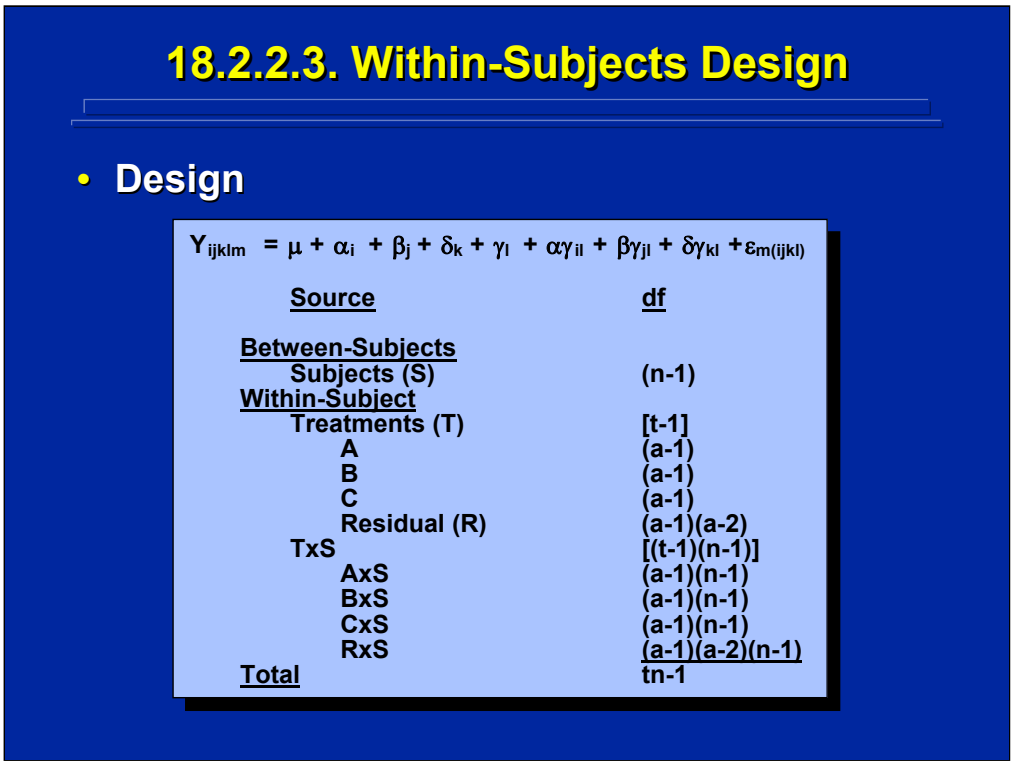
$SS_R = (\sum ABC_{ijk}.^2/n) - (\sum A_{i}...^2/an) - (\sum B._{j}..^2/an) - (\sum C.._{k}.^2/an) + 2(T...^2/tn)$

$SS_{S/T} = \sum ABCS_{ijkl}^2 - (\sum ABC_{ijk}.^2/n)$

$SS_{Total} = \sum ABCS_{ijkl}^2 - (T....^2/tn)$

- **Error Term**: $MS_{S/T}$

The SS computational formulae shown on this slide for between-subjects Latin square designs follow the standard procedures in basic ANOVA with a slight modification for the Residual. The Residual effect is between-cell variation or $SS_{Total}$ adjusted for the presence of the three main effect variations.




The statistical model under the additivity assumption for a within-subjects Latin square design is stated at the top of this slide. This slide also summarizes the sources and degrees of freedom for a within-subjects Latin square design in which each of “n” subjects receives each of the treatment conditions in the Latin square design. The Factor A, B, and C main effects and the Residual (R) effect are merely listed under treatments, because the sum of squares of these sources add to the total treatment effect (T). Likewise, the interaction of these effects with subjects are listed under the TxS. Standard within-subject error terms can be used in F-tests of Factors A,B, C, and R. Namely, each of these effects is tested by its interaction with subjects (e.g., AxS is the error term for A).

## 18.2.2.3. Within-Subject Designs (Cont'd)

- **Sum of Squares Computational Formulae**

$$SS_S = (\sum S\ldots_{\ l}{}^2/a^2) - (T\ldots.\ {}^2/tn)$$
$$SS_T = (\sum ABC_{ijk.}{}^2/n) - (T\ldots.\ {}^2/tn)$$
$$SS_A = (\sum A_{i\ldots}{}^2/an) - (T\ldots.\ {}^2/tn)$$
$$SS_B = (\sum B._{j..}{}^2/an) - (T\ldots.\ {}^2/tn)$$
$$SS_C = (\sum C.._{k.}{}^2/an) - (T\ldots.\ {}^2/tn)$$
$$SS_R = (\sum ABC_{ijk.}{}^2/n) - (\sum A_{i\ldots}{}^2/an) - (\sum B._{j..}{}^2/an) - (\sum C.._{k.}{}^2/an) + 2(T\ldots\ {}^2/tn)$$
$$SS_{TxS} = \sum ABCS_{ijkl}{}^2 - (\sum ABC_{ijk.}{}^2/n) - (\sum S\ldots_{\ l}{}^2/a^2) + (T\ldots.\ {}^2/tn)$$
$$SS_{AxS} = (\sum AS_{i..l}{}^2/a) - (\sum A_{i\ldots}{}^2/an) - (\sum S\ldots_{\ l}{}^2/a^2) + (T\ldots.\ {}^2/tn)$$
$$SS_{BxS} = (\sum BS._{j.l}{}^2/a) - (\sum B._{j..}{}^2/an) - (\sum S\ldots_{\ l}{}^2/a^2) + (T\ldots.\ {}^2/tn)$$
$$SS_{CxS} = (\sum CS.._{kl}{}^2/a) - (\sum C.._{k.}{}^2/an) - (\sum S\ldots_{\ l}{}^2/a^2) + (T\ldots.\ {}^2/tn)$$
$$SS_{RxS} = SS_{TxS} - SS_{AxS} - SS_{BxS} - SS_{CxS}$$
$$SS_{Total} = \sum ABCS_{ijkl}{}^2 - (T\ldots.\ {}^2/tn)$$

- **Error Terms**: Interactions With Subjects

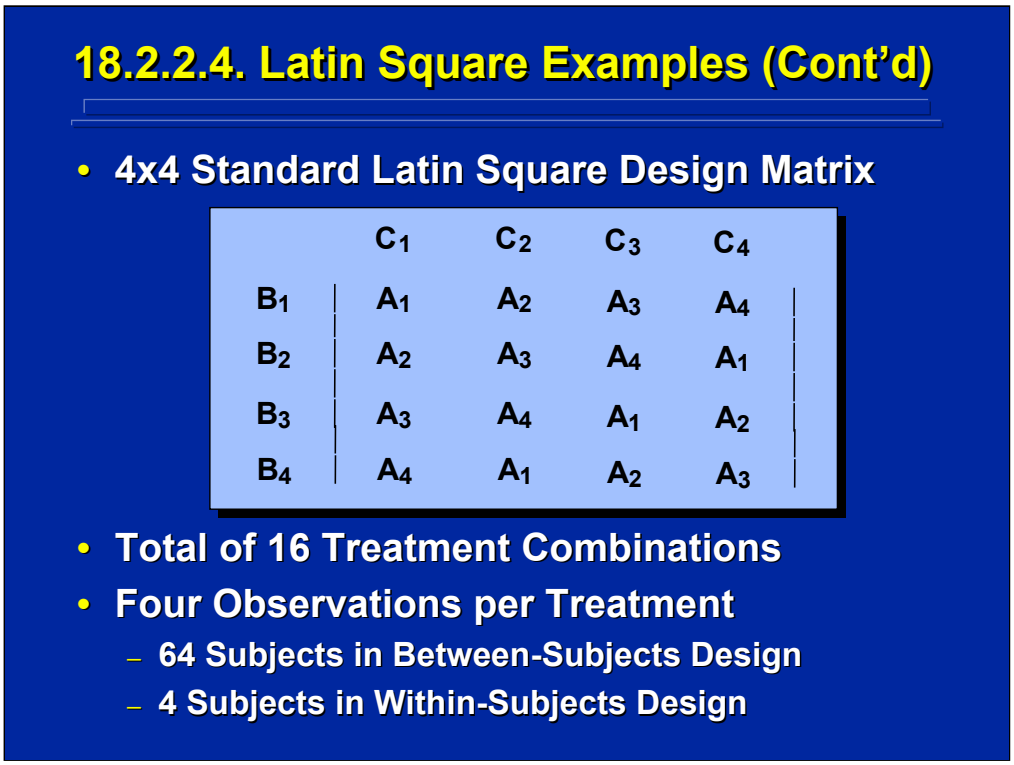
The SS computational formulae summarized on this page for within-subjects Latin square designs follow the standard procedures in basic ANOVA with a slight modification for the Residual effects, $SS_R$ and $SS_{RxS.}$ These Residual effects correct the overall treatment effect for the presence of the three main effects, A, B, and C.

## 18.2.2.4. Latin Square Examples

- **Example Problem. The main effects of three characteristics of a hand held communication device was evaluated by forward observers in Army training exercises. Four different levels each of Input Display Color Resolution (A), Speaker Characteristics (B), and Keys Size (C), of the devices were evaluated in a 4x4 standard Latin square design. The minutes to complete a communication were measured on four soldiers in each treatment combination. Did any of the three characteristics of the communication devices have a significant effect on time to communicate ($p < 0.01$)?**

(Click in this red rectangle to see SAS calculations for this example.)

This example describes a 4x4 standard Latin square design used to evaluate four levels each of three factors in the interface design of a hand held communication device. Since the experimenter is only interested in main effects, a Latin square design is appropriate. This evaluation can be conducted using either a between-subjects or a within-subjects design depending on how subjects are assigned to treatment conditions. Consequently, both solutions for this example problem are provided. The Slater and Williges (2006) appendix provides the SAS program for conducting the ANOVA on both the between-subjects and the within-subjects versions of this Latin square design.




(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the 16 treatment combinations in a 4x4 standard Latin square design that is designated by the four levels of Factor C as columns and the four levels of Factor B as rows of the square. The four levels of Factor A are listed in numerical order in the first row and column of the Latin square. The square is completed by adding one to each level of the first column of Factor A such that each level of Factor A appears once in each row and column of the standard Latin square design.

This 4x4 Latin square defines the 16 combinations of the four display color resolutions (Factor A), the four key sizes used for input (Factor B), and the four types of speaker (Factor C) used in the hand held communication devices being evaluated. This design is replicated four times to yield four observations in each of the 16 cells of the design. If the design is a between-subjects design, each of the 16 hand held communication devices is used by a different group of four soldiers requiring a total of 64 soldiers to complete the experiment. If the design is a within-subjects design, the same four soldiers use all 16 versions of the hand held communication device. Each soldier in the within-subjects design alternative would use the 16 hand held communication devices in a random order to minimize treatment order bias.

## 18.2.2.4. Latin Square Examples (Cont'd)

- **Data Matrix for 4x4 Latin Square Examples**

| $A_1B_1C_1$ | $A_2B_1C_2$ | $A_3B_1C_3$ | $A_4B_1C_4$ | $A_2B_2C_1$ | $A_3B_2C_2$ | $A_4B_2C_3$ | $A_1B_2C_4$ |
|---|---|---|---|---|---|---|---|
| 15 | 25 | 30 | 39 | 25 | 35 | 40 | 26 |
| 20 | 26 | 32 | 33 | 37 | 42 | 49 | 35 |
| 22 | 30 | 28 | 28 | 39 | 39 | 42 | 32 |
| 18 | 25 | 32 | 35 | 28 | 40 | 38 | 28 |
| $A_3B_3C_1$ | $A_4B_3C_2$ | $A_1B_3C_3$ | $A_2B_3C_4$ | $A_4B_4C_1$ | $A_1B_4C_2$ | $A_2B_4C_3$ | $A_3B_4C_4$ |
| 30 | 45 | 28 | 38 | 39 | 21 | 15 | 20 |
| 36 | 47 | 35 | 35 | 35 | 30 | 24 | 22 |
| 28 | 44 | 30 | 36 | 38 | 25 | 18 | 15 |
| 32 | 40 | 29 | 38 | 40 | 22 | 30 | 25 |

(Click in this red rectangle to see SAS calculations for this example.)

This slide summarizes the results of four replications of the 16 treatment conditions in the 4x4 Latin square design. Each of the four numbers shown under each treatment combination is the time required to complete the communication task in minutes using a particular hand held communication device defined by a combination of the levels of Factors A, B, and C. The combination of levels for each hand held communication device represents the 16 treatment conditions defined by the 4x4 standard Latin square design shown on the previous slide.

If a between-subjects Latin square design was used, the 64 communication task completion times shown on this slide represent 64 different soldiers. If a within-subjects Latin square design was used, then the first number under each of the 16 treatment combinations is time required by soldier 1 to complete the communication task, followed by the listing of completion times required by soldiers 2, then 3, then 4.

## 18.2.2.4. Latin Square Examples (Cont'd)

- **Between-Subjects 4x4 Latin Square Design**
  - **ANOVA Summary Table**

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Treatments (T) | [15] | [3436.11] | [229.07] | [14.75*] |
| Color Resolution (A) | 3 | 1602.17 | 534.06 | 34.40* |
| Speakers (B) | 3 | 1316.80 | 438.93 | 28.27* |
| Input Keys (C) | 3 | 115.17 | 38.39 | 2.47 |
| Residual (R) | 6 | 401.97 | 66.99 | 4.31* |
| S/T | 48 | 745.25 | 15.52 | |
| Total | 63 | 4181.36 | | *p < 0.01 |

(Click in this red rectangle to see SAS calculations for this example.)

This slide depicts the ANOVA Summary Table for the analysis of the data shown on the previous page for a between-subjects alternative of the 4x4 standard Latin square design used in data collection. The Slater and Williges (2006) appendix provides the SAS solution for this between-subjects design. The abbreviations for the three independent variables are listed as A, B, and C to make them compatible with the previous slides for this example. Normally, a meaningful abbreviation is chosen for each factor.

The S/T effect is used as the error term for each F-test. The main effects of display color resolution and type of speaker has a significant effect on communication completion time. Post hoc analyses are required to isolate these significant effects. Since the residual effect is significant, some type of interaction exists among the three main effects and this source of variation cannot be combined with S/T to form a pooled error term. A follow-on complete factorial design is needed to isolate significant interaction(s) present in this experiment.

## 18.2.2.4. Latin Square Examples (Cont'd)

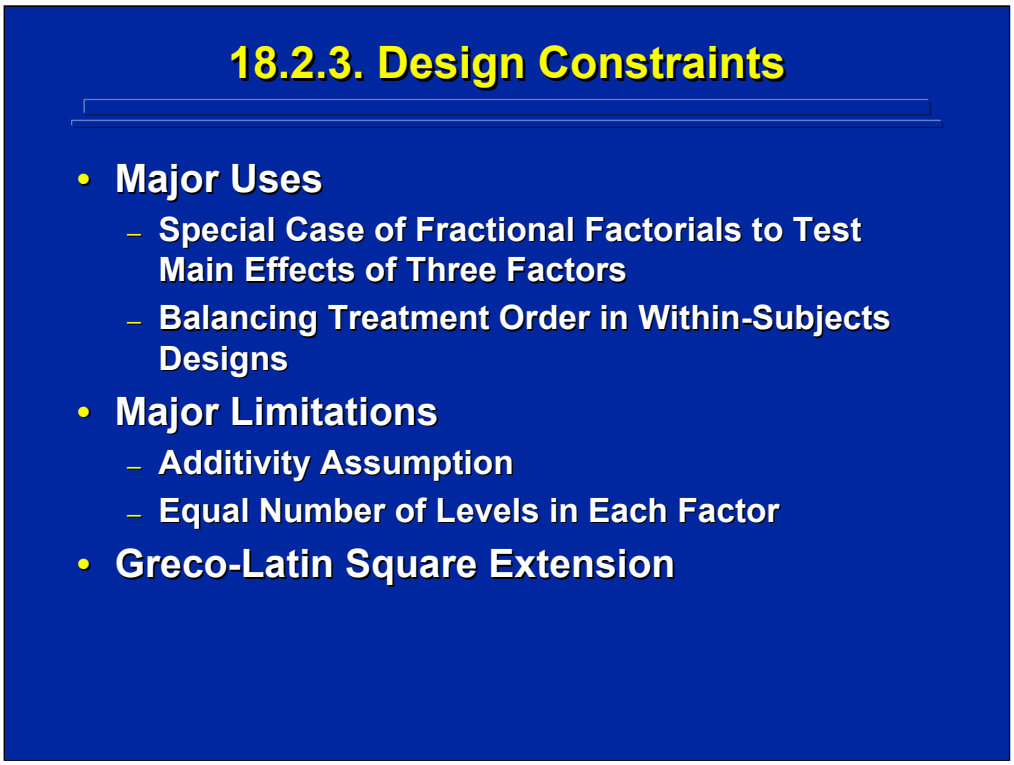
- **Within-Subjects 4x4 Latin Square Design**
  - **ANOVA Summary Table**

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Between-Subjects | | | | |
| Subjects (S) | 3 | 144.92 | 48.31 | |
| Within-Subject | | | | |
| Treatments (T) | [15] | [3436.11] | [229.07] | [17.17*] |
| Color Resolution (A) | 3 | 1602.17 | 534.06 | 28.19* |
| Speakers (B) | 3 | 1316.80 | 438.92 | 20.27* |
| Input Keys (C) | 3 | 115.17 | 38.39 | 2.84 |
| Residual (R) | 6 | 401.97 | 66.99 | 10.63* |
| TxS | [45] | [600.33] | [13.34] | |
| AxS | 9 | 170.52 | 18.95 | |
| BxS | 9 | 194.89 | 21.65 | |
| CxS | 9 | 121.52 | 13.50 | |
| RxS | 18 | 113.40 | 6.30 | |
| Total | 63 | 4181.36 | | *p < 0.01 |

(Click in this red rectangle to see SAS calculations for this example.)

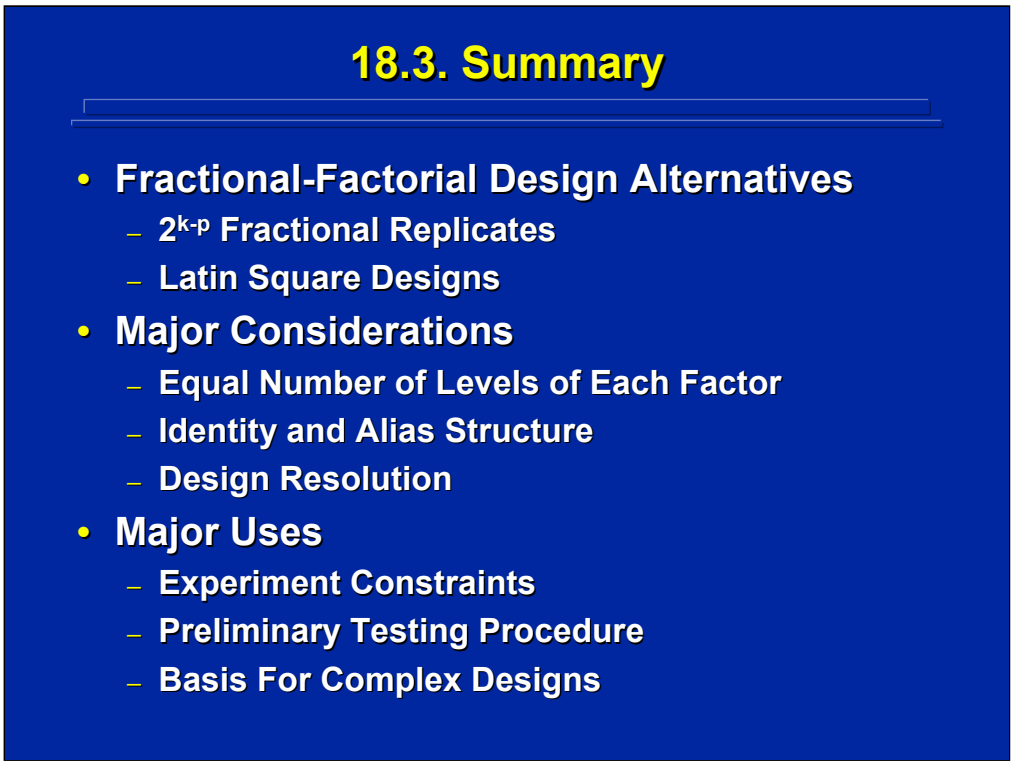
The slide shows the ANOVA Summary Table of the within-subjects design alternative of the 4x4 Latin square design using the example problem results. The Slater and Williges (2006) appendix provides the SAS solution for this within-subjects design. Again, the abbreviations for the three independent variables are listed as A, B, and C to make them compatible with the previous slides for this example. Normally, a meaningful abbreviation is chosen for each factor.

The interaction of Subjects (S) with treatment effects is used as the error term to test each effect. The main effects of display color resolution and type of speaker has a significant effect on communication completion time. Post hoc analyses are required to isolate these significant effects. Since the residual effect is significant, some type of interaction exists among the three main effects, and this source of variation cannot be combined with TxS to form a pooled error term. A follow-on complete factorial design is needed to isolate significant interaction(s) present in this experiment.




Latin square designs can be used as a special case of between-subjects and within-subjects, fractional-factorial designs to test only the main effects of three factors. In human factors research, balanced Latin square designs are also used to partially counterbalance the treatment order in within-subjects designs as described in Topic 12. The experimenter must assume additivity, or no interaction among the three factors, in order to construct unbiased F-tests. In addition, an equal number of levels of each of the three factors is required to construct the Latin square design.

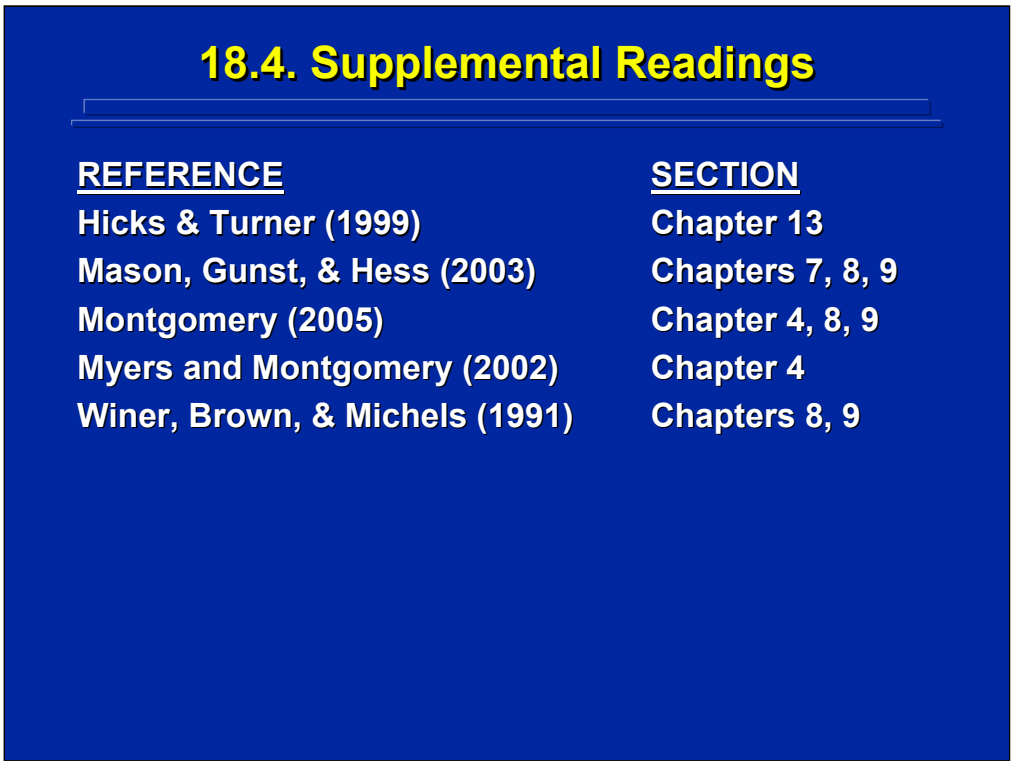
The basic Latin square design can be extended to consider more than three factors. Greco-Latin square designs consider four factors by combining two orthogonal Latin squares. This could be extended beyond four factors through hyper-Greco-Latin squares. However, there are only a limited number of these design alternatives due to the requirement for orthogonal Latin square components. Winer, et al. (1991, Chapter 9) provide a description of Greco-Latin square design and analysis procedures.




Fractional-factorial designs are used in human factors and ergonomics research when only a fractional component of the full factorial design is investigated. One-half and one-fourth replicates of $2^k$ factorials are the most often used fractional-factorial designs due to their straightforward confounding structure. When the experimenter is only interested in investigating main effects of three factors and the levels of each factor are equal, Latin square designs can be considered as the fractional replicate.

The experimenter must consider three major components of fractional factorials carefully. First, these designs require that each factor is observed at the same number of levels. Second, some information in the full factorial design has to be sacrificed in the fractional factorial. The identity relationship effect is completely lost and the alias structure specifies the confounded effects. Third, the experimenter must choose the appropriate design resolution to insure orthogonal evaluation of main effects and interactions of interest in the fractional replicate.

Fractional-factorial designs are useful in human factors research when time, equipment, and budgets constraints preclude use of complete factorial designs. In large experiments, a fractional replicate may provide an efficient method of pre-testing. Finally, fractional-factorial designs form components of complex designs used in empirical model building and sequential experimentation as discussed in Section 5.

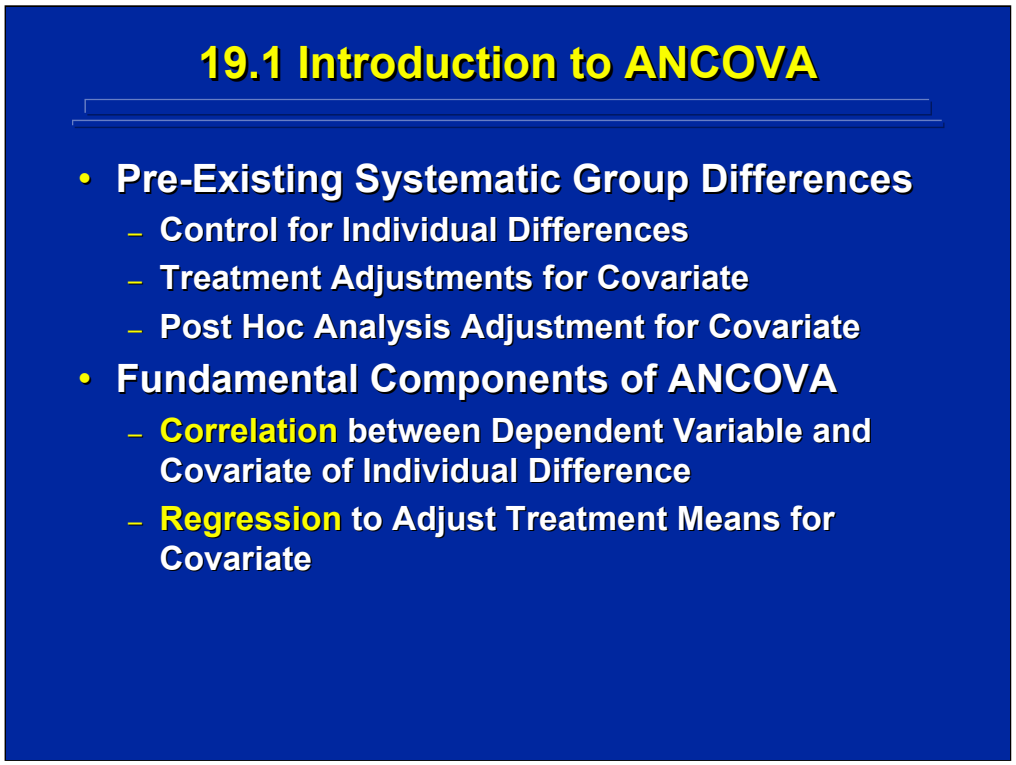



All these texts provide a discussion of fractional-factorial and Latin square designs used in ANOVA. Winer, et al. (1991) provide a detailed description of the modular representation approach used in this topic to construct fractional-factorial ANOVA designs and provide a complete description of design construction, analysis, and alternatives of Latin square designs used in behavioral research.

# Topic 19. Analysis of Covariance (ANCOVA)

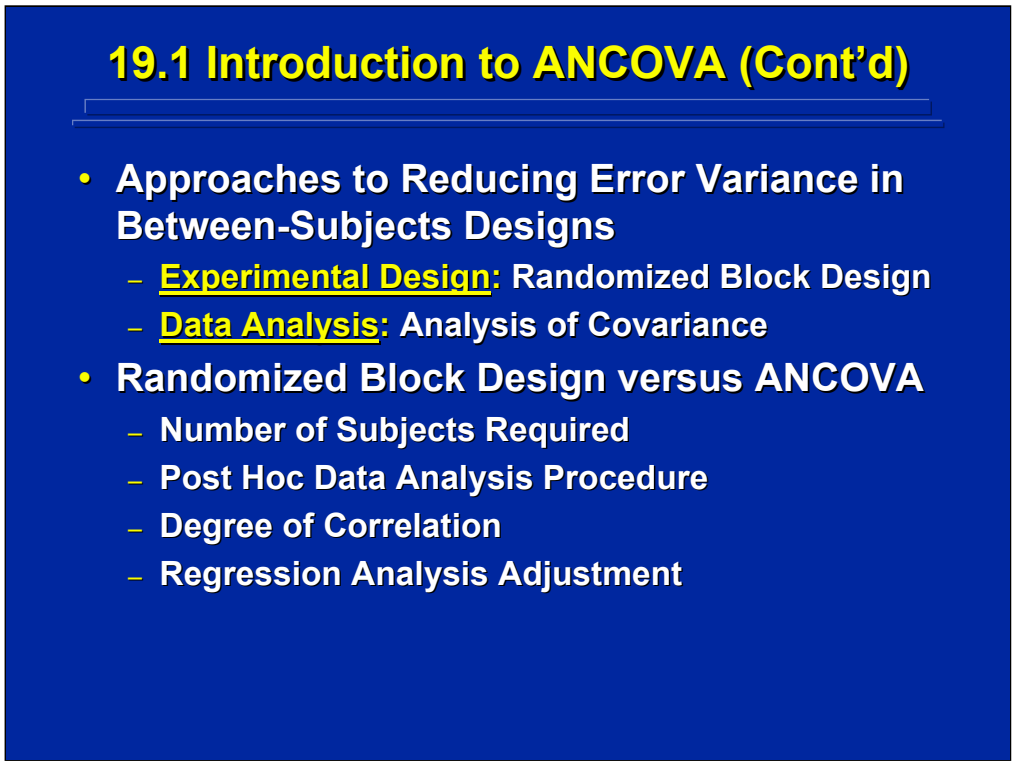
19.1. Introduction to ANCOVA
19.2. Linear Correlation
19.2.1. Correlation Coefficient, $r_{12}$
19.2.2. Alternative Correlations
19.3. Simple Linear Regression
19.3.1. Line of Best Fit
19.3.2. Goodness of Fit
19.4. ANCOVA Computations
19.4.1. Basic ANCOVA Design
19.4.2. Advanced ANCOVA
19.4.3. Interpreting ANCOVA
19.5. Summary
19.6. Supplemental Readings

This topic deals with an analytical technique for reducing the effect of a covariate to increase the sensitivity of the F-test on effects of interest to the experiment. The covariate is correlated with the dependent variable, and its effect is removed through simple linear regression. Consequently, both calculations of correlation and simple regression are described as the basic components of analysis of covariance (ANCOVA). Basic computations in ANCOVA and subsequent interpretations of results are described in this topic. Supplemental readings on correlation, simple regression, and ANCOVA are provided.




ANCOVA is a well-accepted analytical procedure used to adjust for pre-existing systematic differences between groups. For example, in training research, different training methods may be evaluated in different classes such that each class receives a different training method. But, the students in one class may differ greatly in terms of verbal abilities of students in another class. ANCOVA can be used to adjust the various training groups for individual differences in class verbal aptitudes that are correlated with learning based on the different training procedures tested. Both the main ANCOVA analysis and post hoc analyses on significant effects are adjusted for the verbal abilities covariate to provide a more sensitive test of the training methods.

The ANCOVA procedure is based on the correlation between an individual difference component and the dependent variable. Regression procedures are then used to remove the covariate effect and a subsequent ANOVA is conducted on the adjusted treatment means.



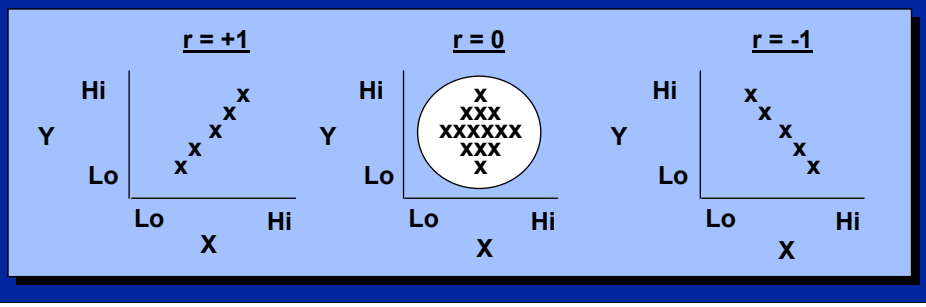

ANCOVA is also a statistical procedure for refining error variance in a between-subjects design to provide a more sensitive F-test. The randomized block as discussed in Topic 15 is an experimental design alternative to the ANCOVA analytical approach. Randomized block designs control the effect of the covariate through experimental design; whereas, ANCOVA adjusts for the covariate effect statistically.

Both alternatives also have drawbacks. In the randomized block design, the experimenter usually needs to pretest more subjects than required to obtain equal sample sizes for the various levels of the covariate. In ANCOVA, interpretations are made on treatment means adjusted for the covariate rather than the actual treatment means. Randomized block designs are often preferred because no adjustment to the means and subsequent interpretations are required.

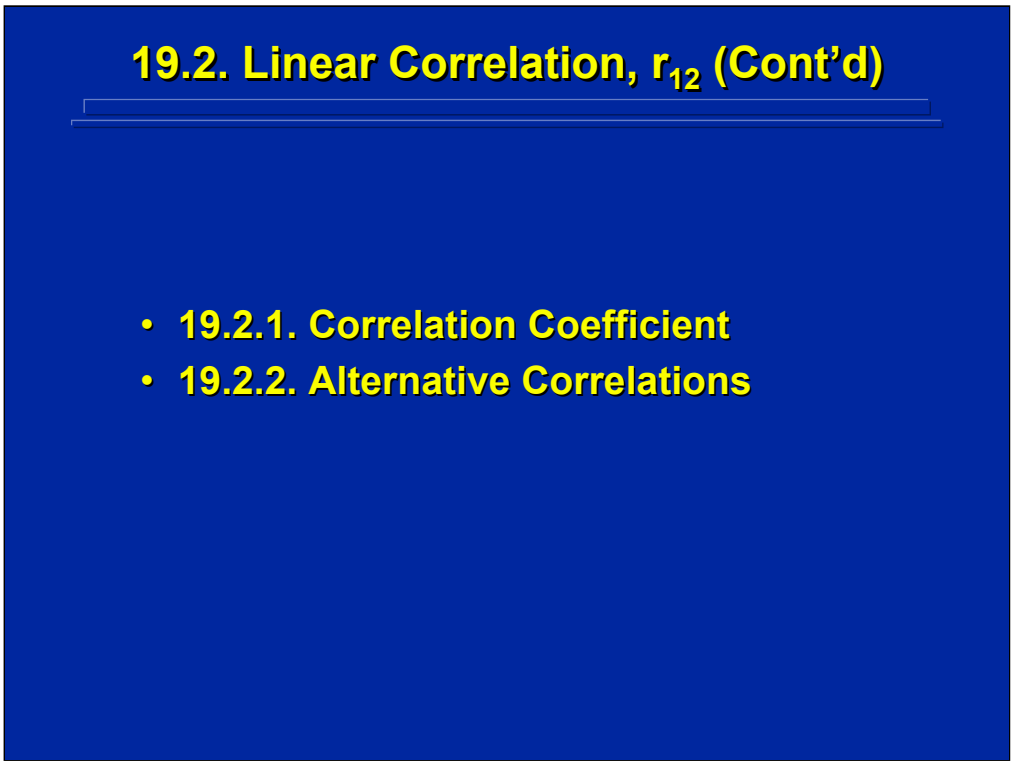
Both approaches take into account the effect of a covariate that is correlated with the dependent variable. To conduct the ANCOVA, both linear correlation and regression procedures need to be used for adjusting and analyzing the treatments of interest. Consequently, both the concepts of linear correlation and simple linear regression are reviewed in the next two subsections as a precursor to ANCOVA computations.

## 19.2. Linear Correlation, $r_{12}$

- **Definition**: Quantitative description of the degree of linear relationship between two variables.
  - Pearson-Product Moment Correlation, $r_{12}$
  - Linear Relationship
  - Range: r = +1 to –1
  - Scatterplot




Correlation designated by $r_{12}$ is the description of the linear relationship between two variables. If there is a perfect positive linear relationship, then $r_{12}$ = 1. If there is a perfect negative linear relationship, $r_{12}$ = -1. If $r_{12}$ = 0, then there is no linear relationship between the two variables and the scatter plot between the two variables is circular as shown in the middle diagram on this slide. Consequently, the linear correlation between two variables ranges somewhere between +1 and -1.




Although the Pearson product-moment correlation coefficient, $r_{12}$, is the primary measure of linear correlation between two variables, various alternatives to $r_{12}$ are available to handle special circumstances. Formulae for both the Pearson $r_{12}$ and some of its alternatives are presented in this subsection.

## 19.2.1. Correlation Coefficient

- **Pearson Product-Moment Correlation, $r_{12}$**
  - **Definitional Form**

$$r_{12} = \frac{\sum z_x z_y}{n-1}$$

$$= \frac{\sum \left(\frac{X-\bar{X}}{s_X}\right)\left(\frac{Y-\bar{Y}}{s_Y}\right)}{n-1}$$

$$r_{12} = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{(n-1)s_X s_Y}$$

**Note:** $(X-\bar{X})$ and $(Y-\bar{Y})$ = **First Moment**

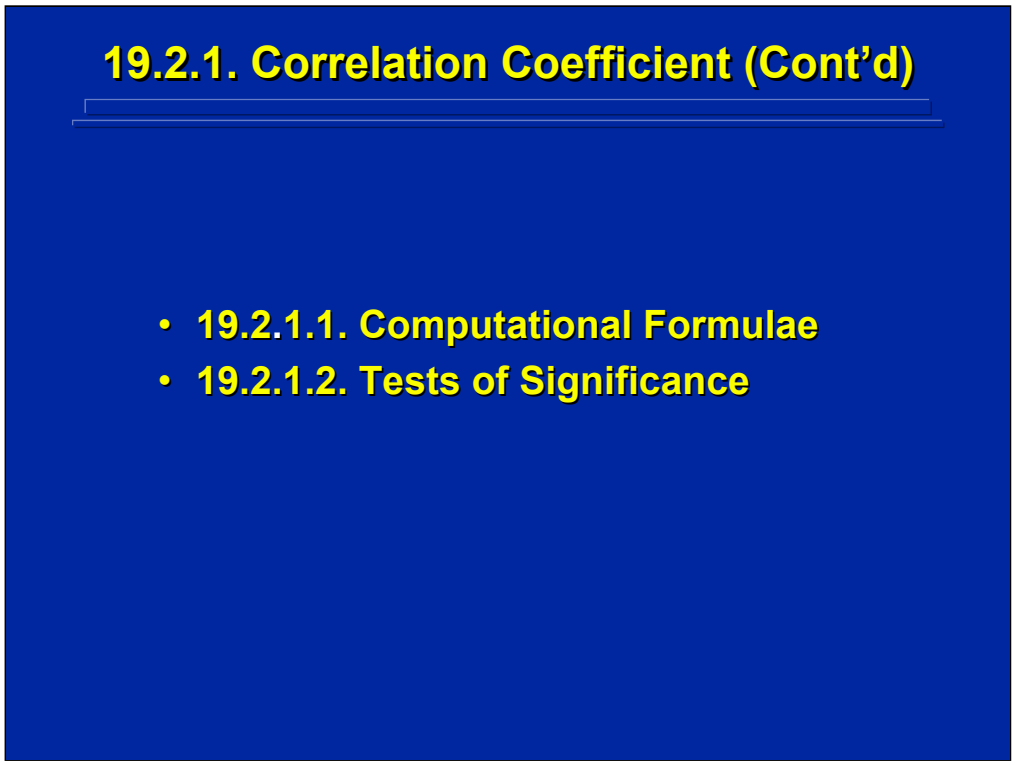
This slide shows the definitional formula for the Pearson $r_{12}$. Pearson defined the sum of the product of Z scores for two variables, X and Y, divided by n-1 degrees of freedom as the correlation $r_{12}$. The Z scores are standardized scores are defined in Topic 3 on page 89 in this reference material. Note that the numerator of the final formula for $r_{12}$ shown on this slide is also the sum of the product of the first moment around the X and Y means, respectively. Hence the name of $r_{12}$ is the Pearson product-moment correlation coefficient.

## 19.2.1. Correlation Coefficient (Cont'd)

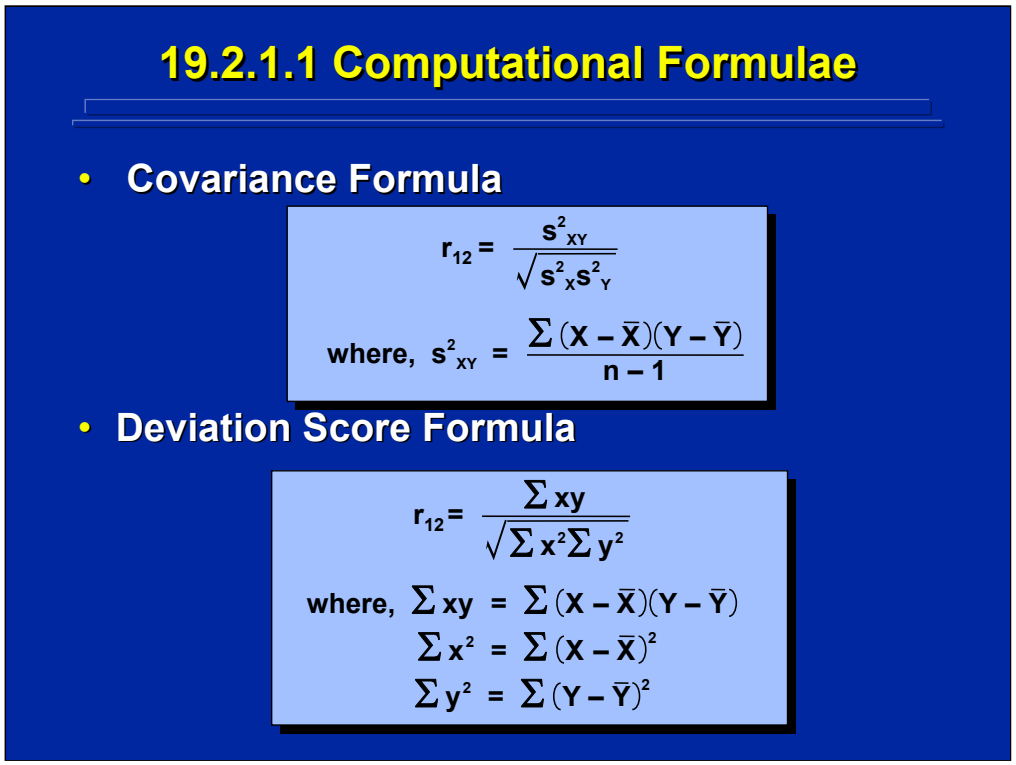
- Interpretation of Pearson $r_{12}$
  - Degree of Linear Relationship
  - Not Percent of Variation
- Not Necessarily Causative
  - Correlation with Third Variable
- Choice of Variables
  - Dependent, Y, and Independent, X, Variables
  - Intercorrelation Matrix
    - Dependent Variables, Y's
    - Independent Variables, X's
  - Prediction Via Regression
    - Y Predicted by X

The correlation describes the linear relationship between two variables somewhere between ±1 and not the percent of variation between two variables, X and Y. The correlation value expresses only the degree of linearity and is not necessarily a causative relationship because both of the two variables correlated could be correlated with a third variable that represents the true causative relationship. Consequently, causative interpretations of correlations should be considered carefully.

Several types of correlations between two variables are used in experimental design. A dependent variable is designated Y, and an independent variable is designated X. An intercorrelation matrix can be calculated among several dependent variable, Y's, or several independent variables, X's. Correlations can be used in regression to predict Y as a function of one or more X's. Instead of doing hypothesis testing, the experimenter may want to build an empirical model where Y is predicted by X's as described in Section 5.




This subsection describes various computational formulae and tests of significance for the basic Pearson product-moment correlation, $r_{12}$.




This slide shows the covariance or deviation score formulae for calculating the Pearson product-moment correlation, $r_{12}$. Note that the covariance between X and Y is shown in the numerator of the covariance formulae at the top of the slide. Deviation scores are the differences between a score and its mean. They are listed as lowercase x and y letters and are defined in terms of $\Sigma xy$, $\Sigma x^2$, and $\Sigma y^2$ in the lower portion of this slide. The $\Sigma xy$ value is the sum of cross products of the X and Y deviations, the $\Sigma x^2$ value is the sum of squared deviations of X, and the $\Sigma y^2$ value is the sum of squared deviations of Y.

## 19.2.1.1 Computational Formulae (Cont'd)

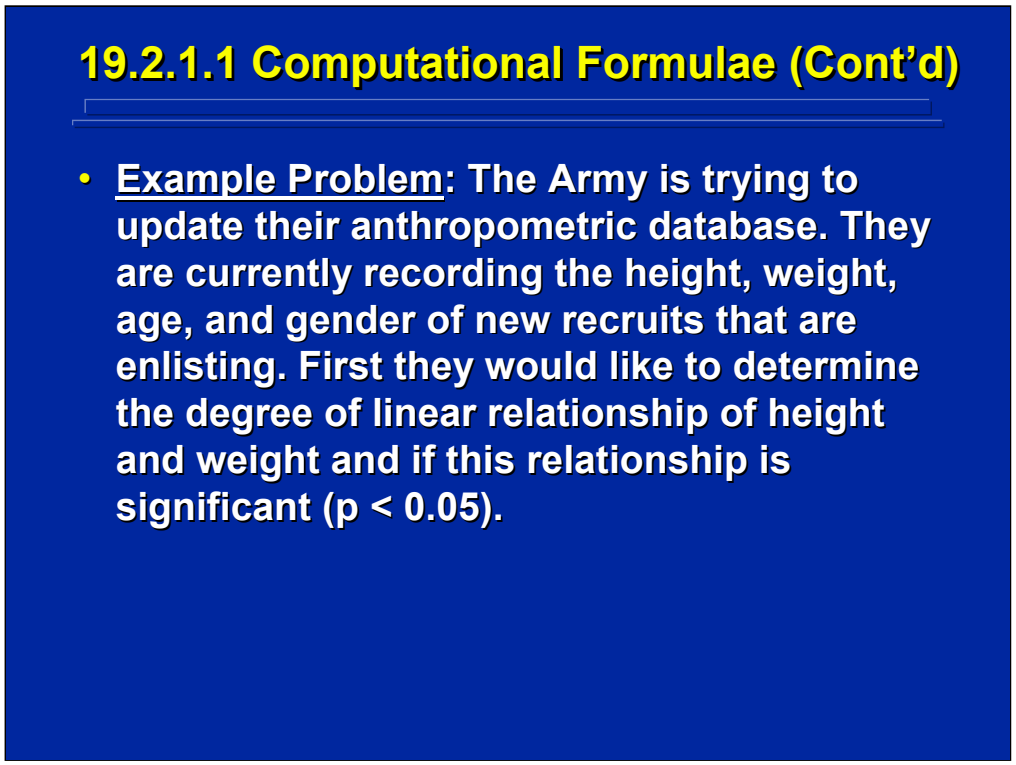
- **Raw Score Formula**

$$r_{12} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}}$$

– or –

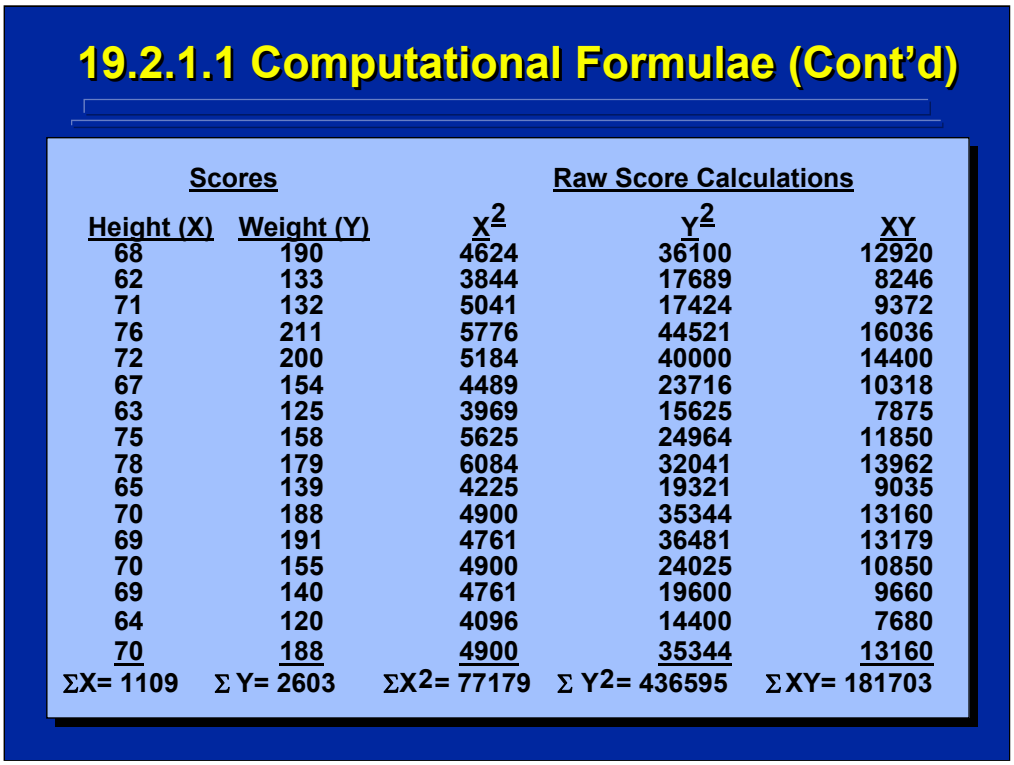
$$r_{12} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right]\left[n\sum Y^2 - (\sum Y)^2\right]}}$$

The two raw score formulae for $r_{12}$ shown on this slide use no intermediate mean calculations. Both are algebraically equivalent, but the formula shown on the bottom of this slide is the most common version of the Pearson product-moment correlation coefficient.




(Click in this red rectangle to see SAS calculations for this example.)

This example problem lists four variables (i.e., height, weight, age, and gender) that can be correlated to show the linear relationship between any two of them. Specifically, this problem asks for the value of linear correlation between height (X) and weight (Y) and if this correlation is statistically significant ($p < 0.05$).




(Click in this red rectangle to see SAS calculations for this example.)

This slide presents hypothetical anthropometric data in terms of height (X) and weight (Y) of 16 soldiers. The raw score values for the sum of X and Y, the sum of squares of X and Y, and the sum of cross products XY are shown on this slide. The appendix by Slater and Williges (2006) provides the SAS program solutions for the various correlation examples provided in this reference material.

## 19.2.1.1 Computational Formulae (Cont'd)

**Raw Score Calculations**

$$r_{12} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{\left[n\Sigma X^2 - (\Sigma X)^2\right]\left[n\Sigma Y^2 - (\Sigma Y)^2\right]}}$$

$$r_{12} = \frac{(16)\ (181703)\ -\ (1109)\ (2603)}{\sqrt{\left[(16)\ (77179) - (1109)^2\right]\left[(16)\ (436595) - (2603)^2\right]}} = +0.635$$

(Click in this red rectangle to see SAS calculations for this example.)

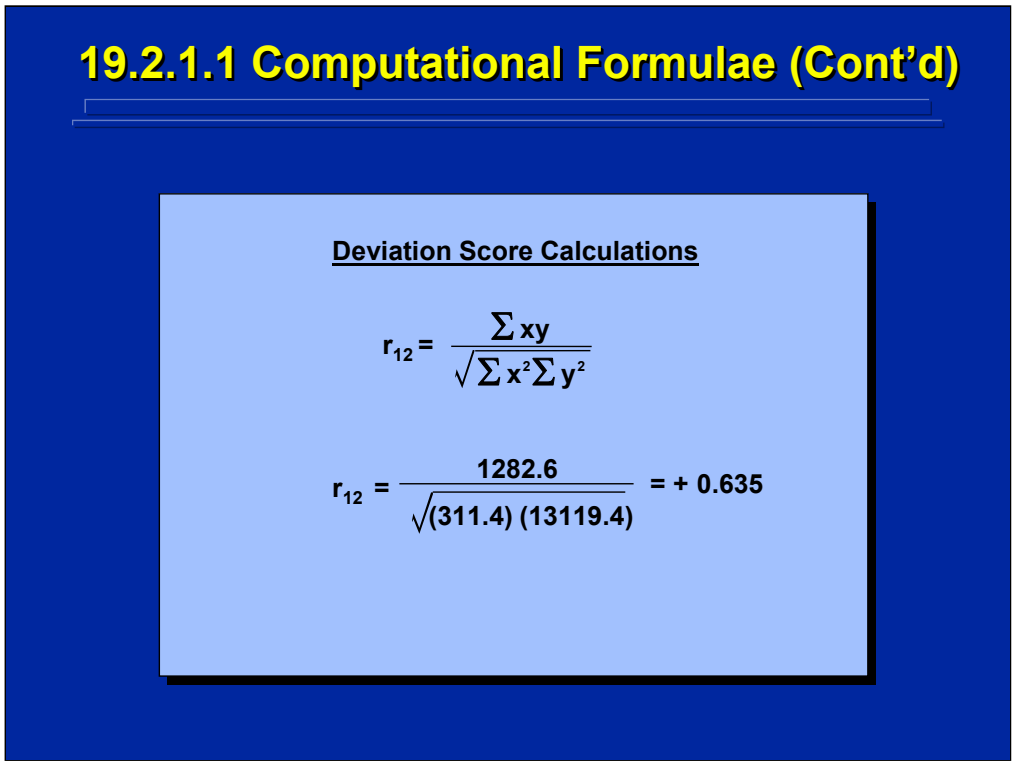
The correlation between the height and weight data of the 16 soldiers shown on the previous slide is calculated using the raw score formula for $r_{12}$. The resulting correlation is +0.635. This shows a positive linear relationship between height and weight or that weight increases as height increases.

## 19.2.1.1 Computational Formulae (Cont'd)

| Raw Scores | | Deviation Score Calculations | | | | |
|---|---|---|---|---|---|---|
| Height (X) | Weight (Y) | x | y | $x^2$ | $y^2$ | xy |
| 68 | 190 | - 1.3 | 27.3 | 1.7 | 746.0 | - 35.8 |
| 62 | 133 | - 7.3 | - 29.7 | 53.5 | 881.3 | 217.1 |
| 71 | 132 | 1.7 | - 30.7 | 2.8 | 941.7 | - 51.8 |
| 76 | 211 | 6.7 | 48.3 | 44.7 | 2334.1 | 323.1 |
| 72 | 200 | 2.7 | 37.3 | 7.2 | 1392.2 | 100.3 |
| 67 | 154 | - 2.3 | - 8.7 | 5.3 | 75.7 | 20.1 |
| 63 | 125 | - 6.3 | - 37.7 | 39.8 | 1420.3 | 237.9 |
| 75 | 158 | 5.7 | - 4.7 | 32.3 | 22.0 | - 26.7 |
| 78 | 179 | 8.7 | 16.3 | 75.5 | 266.1 | 141.7 |
| 65 | 139 | - 4.3 | - 23.7 | 18.6 | 561.1 | 102.2 |
| 70 | 188 | 0.7 | 25.3 | 0.5 | 640.7 | 17.4 |
| 69 | 191 | - 0.3 | 28.3 | 0.1 | 801.6 | - 8.8 |
| 70 | 155 | 0.7 | - 7.7 | 0.5 | 59.1 | - 5.3 |
| 69 | 140 | - 0.3 | - 22.7 | 0.1 | 514.7 | 7.1 |
| 64 | 120 | - 5.3 | - 42.7 | 28.2 | 1822.2 | 226.8 |
| 70 | 188 | 0.7 | 25.3 | 0.5 | 640.7 | 17.4 |
| $\Sigma$ X= 1109 | $\Sigma$ Y= 2603 | $\Sigma$ x= 0 | $\Sigma$ y= 0 | $\Sigma x^2 = 311.4$ | $\Sigma y^2 = 13119.4$ | $\Sigma$ xy= 1282.6 |
| $\bar{X}$ = 69.3 | $\bar{Y}$ = 162.7 | | | | | |

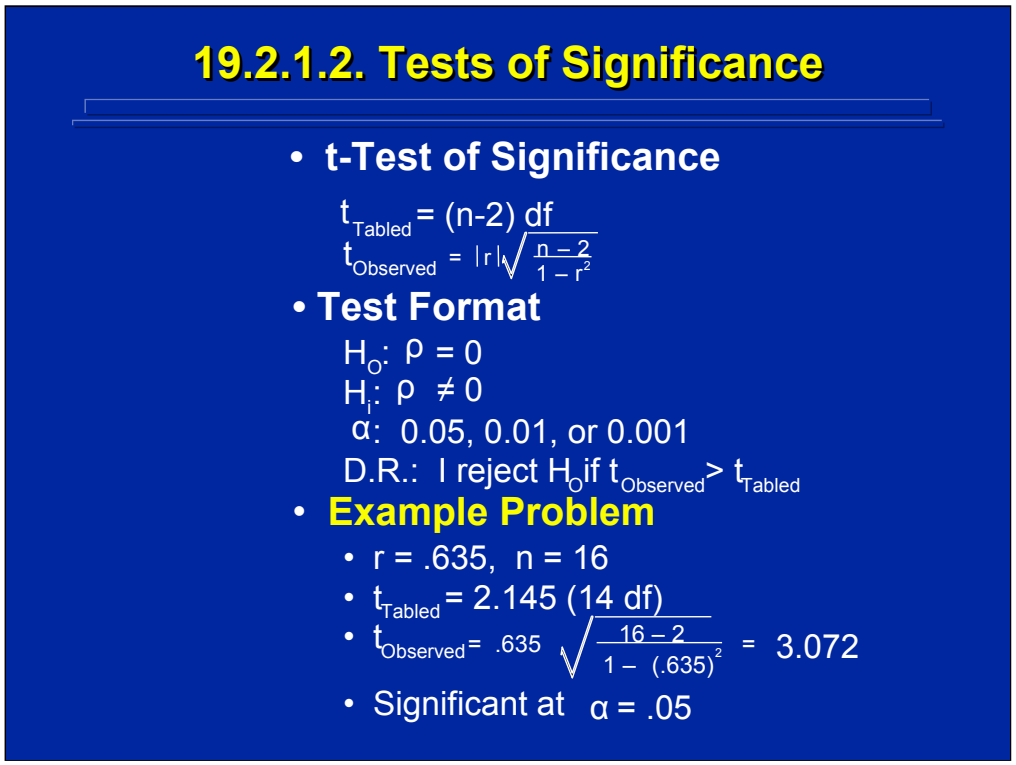
(Click in this red rectangle to see SAS calculations for this example.)

The x, y, and xy deviation scores are listed on this slide for the hypothetical soldier height and weight raw score data and group means for height and weight.




(Click in this red rectangle to see SAS calculations for this example.)

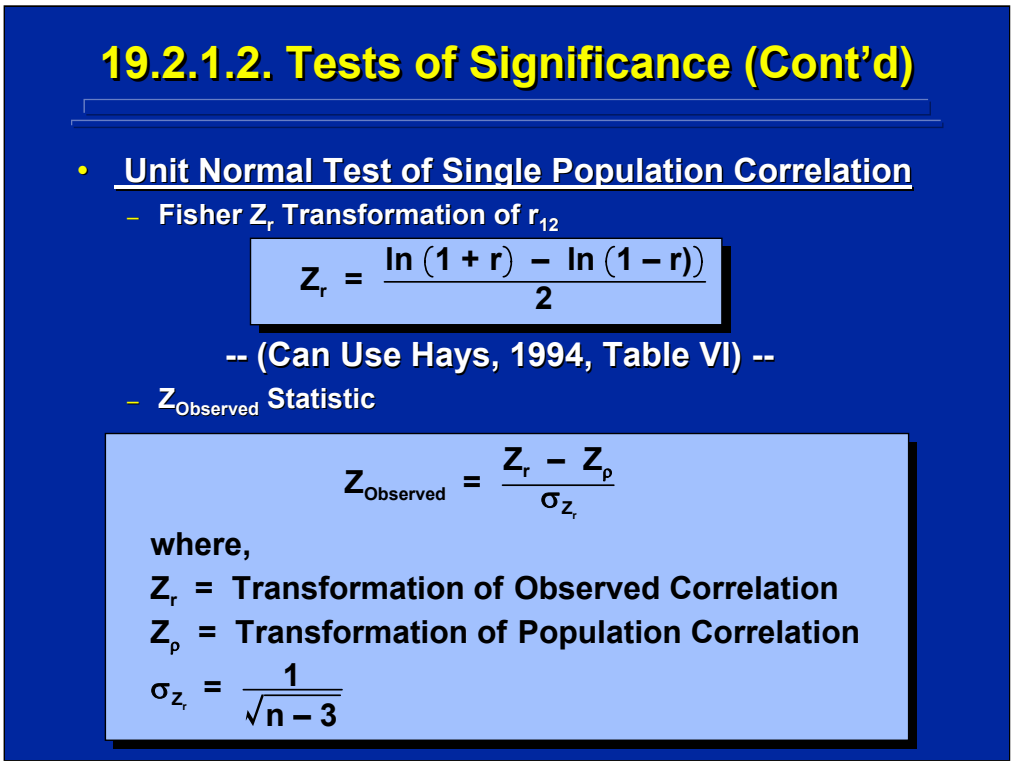
The correlation of the deviation scores shown in the previous slide is calculated using the deviation score formula for $r_{12}$. The resulting correlation between height and weight, +0.635, is the same as the value calculated by the raw score formula for $r_{12}$. Consequently, the formulae are equivalent.




(Click in this red rectangle to see SAS calculations for this example.)

A t-test with n-2 degrees of freedom can be used to determine if a correlation is significantly different from 0. The formula for $t_{Observed}$ is presented at the top of the slide, and the standard test format is provided in the middle of the slide.

The t-test of the example problem correlation of 0.635 is summarized at the bottom on this slide. Since the observed value is 3.072 and the table value is 2.145, the correlation is significant at the 0.05 level.




(Click in this red rectangle to see SAS calculations for this example.)

A second significance test of a correlation is to test it against a known population correlation value. A unit normal test based on the Fisher $Z_r$ transformation can be used to test a correlation against a known population value. The Fisher $Z_r$ transformation formula based on natural logs is provided at the top of the slide. Alternatively, Table VI in Hays (1994) can be used to make the transformation. The resulting formula for the $Z_{Observed}$ value is presented at the bottom of this slide.

## 19.2.1.2. Tests of Significance (Cont'd)

- **Example Problem:** Is a correlation of .635 between height and weight based on a sample size of 16 soldiers significantly different from a population correlation of 0.700 ($p < 0.05$)?
- Test Format

  $H_O$: $\rho = 0.70$

  $H_i$: $\rho \neq 0.70$

  $\alpha$: 0.05

  D.R.: I reject $H_O$ if $Z_{Observed} > Z_{Tabled}$

$$Z_{Observed} = \frac{.7481 - .8673}{.2773} = -0.43 \qquad Z_{Tabled} = -1.96$$

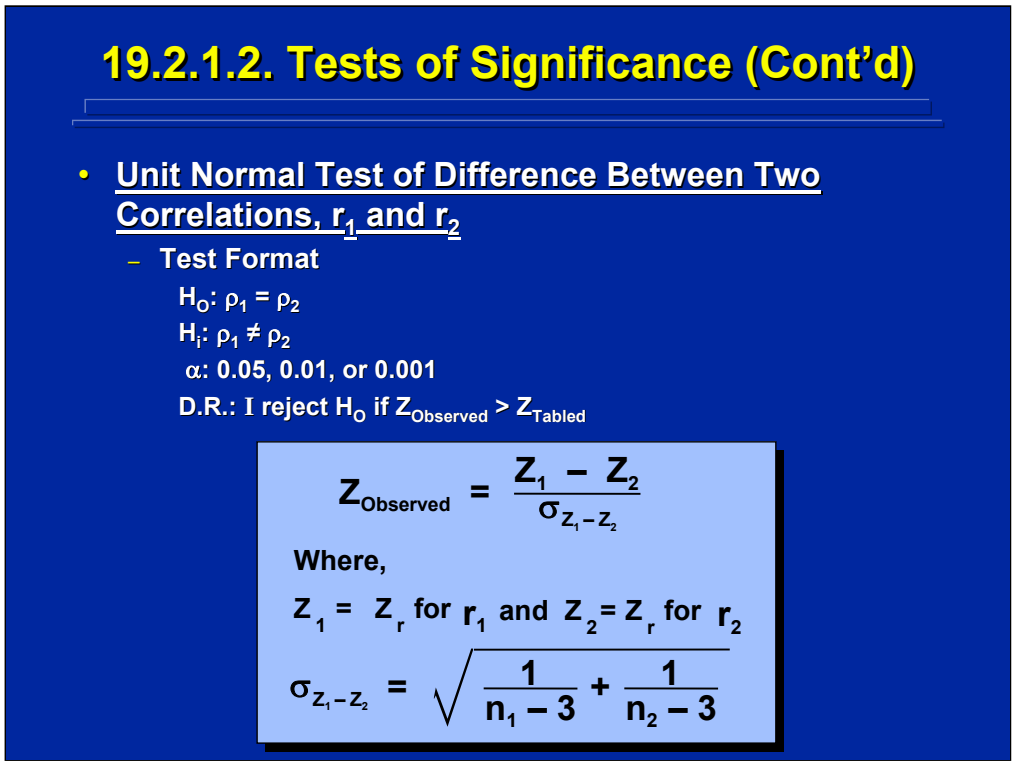
where, $Z_{.635} = .7481$

$Z_{.700} = .8673$

$$\sigma_{Z_r} = \frac{1}{\sqrt{16 - 3}} = .2773$$

(Click in this red rectangle to see SAS calculations for this example.)

In this example, the sample correlation between height and weight of 16 soldiers is compared to a population correlation of 0.700 to test for a significance difference. The test format and $Z_{Observed}$ calculations are shown on this slide. Since the $Z_{Observed}$ value of -0.43 is less that the $Z_{Tabled}$ value of -1.96, the sample correlation is not significantly different ($p < 0.05$) than the population correlation.




(Click in this red rectangle to see SAS calculations for this example.)

The third significance test is to compare two correlations to test for a significant difference between them. Again, the Fisher $Z_r$ transformation can be used to make a Z-test. The test format and the $Z_{Observed}$ formula for this test are shown on the slide.

## 19.2.1.2. Tests of Significance (Cont'd)

- **Example Problem: Is there a significant difference ($p < 0.05$) between the correlations of height and weight for six female, $r_{F(12)} = 0.648$, and six male, $r_{M(12)} = 0.615$, soldiers?**

| Female Soldiers | | Male Soldiers | |
|---|---|---|---|
| Height (X) | Weight (Y) | Height (X) | Weight (Y) |
| 68 | 190 | 78 | 179 |
| 62 | 133 | 65 | 139 |
| 71 | 132 | 70 | 188 |
| 76 | 211 | 69 | 191 |
| 72 | 200 | 70 | 155 |
| 67 | 154 | 69 | 140 |
| 63 | 125 | 64 | 120 |
| 75 | 158 | 70 | 188 |
| $r_{F(12)} = 0.648$ | | $r_{M(12)} = 0.615$ | |

(Click in this red rectangle to see SAS calculations for this example.)

Example data for six female and six male soldiers are shown on this slide. The correlation of height and weight is 0.648 for female soldiers and is 0.615 for male soldiers. Is the difference between these two correlations statistically significant at the 0.05 level?

## 19.2.1.2. Tests of Significance (Cont'd)

- **Test Format**

  $H_O: \rho_1 = \rho_2$

  $H_i: \rho_1 \neq \rho_2$

  $\alpha: 0.05$

  D.R.: I reject $H_O$ if $Z_{Observed} > Z_{Tabled}$

$Z_1 = Z_r$ for 0.648 $\quad Z_2 = Z_r$ for 0.615

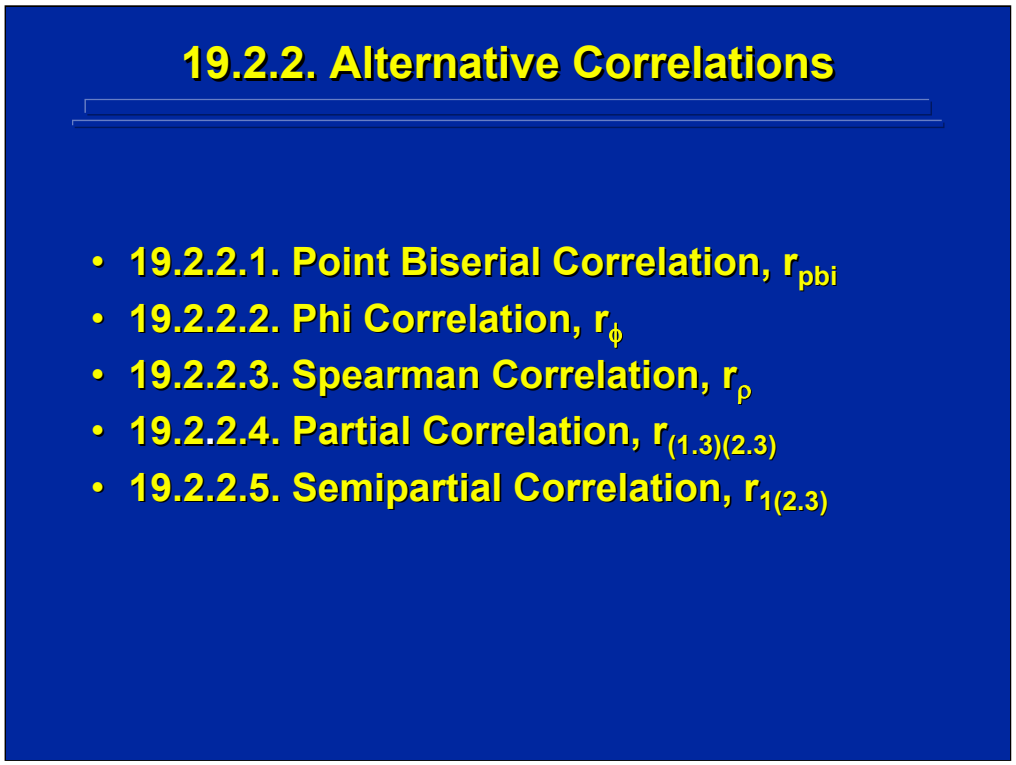
$$\sigma_{Z_1 - Z_2} = \sqrt{\frac{1}{16 - 3} + \frac{1}{16 - 3}} = 0.3922$$

$$Z_{Observed} = \frac{0.7718 - 0.7153}{0.3922} = 0.14$$

$$Z_{Tabled} = 1.96$$

(Click in this red rectangle to see SAS calculations for this example.)

The results of the Z-test on the difference between the two example correlations are shown on this slide. Since the $Z_{Observed}$ value of 0.14 is not greater than the $Z_{Tabled}$ value of 1.96, the experimenter concludes that the correlation of height and weight is not different for female and male soldiers.




Five alternative Pearson correlations are listed on this slide. The first three are nonparametric correlation coefficients. The point biserial coefficient is a correlation of a dichotomous variable with a continuous variable; the phi coefficient is a correlation between two dichotomous variables; and the Spearman rho coefficient is a correlation between two rank orders.

The last alternatives to the Pearson correlation are used when a third variable is considered in the correlation. The partial correlation removes the covariance of the third variable from both of the variables being correlated; whereas, the semipartial correlation removes the covariance of the third variable from only one of the two variables being correlated.

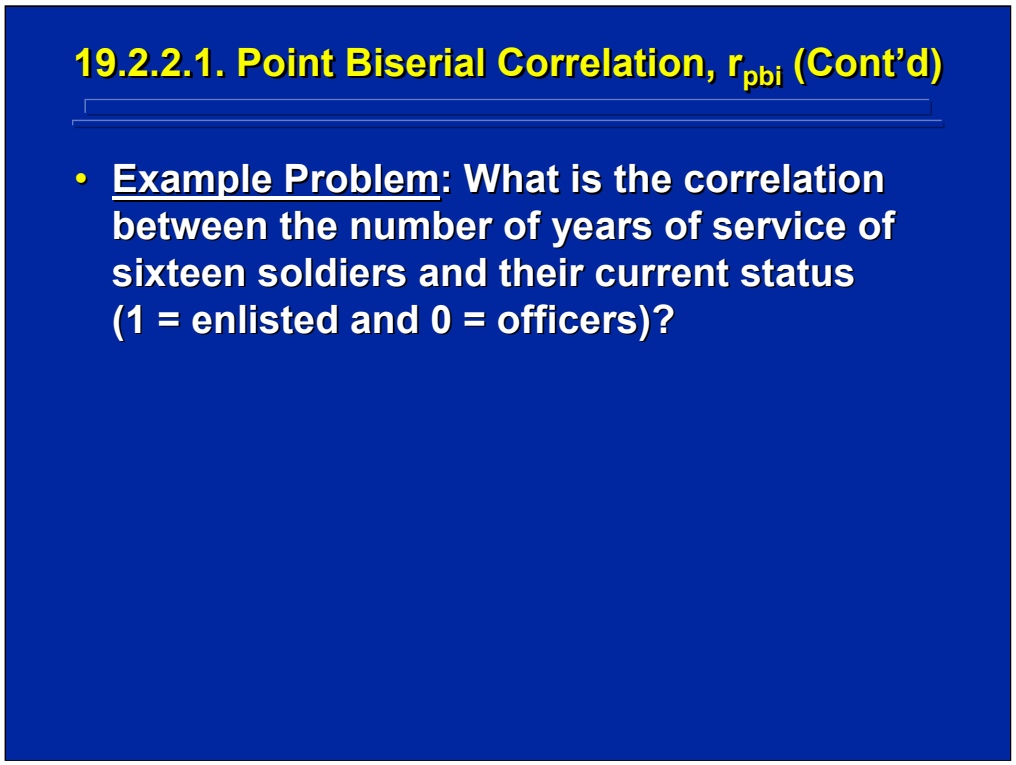
## 19.2.2.1. Point Biserial Correlation, $r_{pbi}$

- **Definition**: Correlation between one continuous and one dichotomous (two category) variable
  - X = Continuous Variable
  - Y = Dichotomous Variable (0, 1)
- Formula

$$r_{pbi} = \frac{\sum X_1 - \frac{n_1 \sum X}{n}}{\sqrt{\frac{n_1 n_0}{n}\left[\sum X^2 - \frac{(\sum X)^2}{n}\right]}}$$

- where,
  $\Sigma X_1$ = sum of X values for n1 observations when Y = 1
  $\Sigma X$ = sum of X values for all n observations
  $n_1$ = number of observations when Y = 1
  $n_0$ = number of observations when Y = 0
  n = total number of observations

This slide shows the formula for the point biserial correlation where X is the continuous variable and Y is the dichotomous variable with only two values, 0 or 1.




(Click in this red rectangle to see SAS calculations for this example.)

This is an example of a problem using the Army anthropometric data on a total of sixteen soldiers that requires a point biserial correlation between number of years of service (continuous) and their current status (dichotomous).

## 19.2.2.1. Point Biserial Correlation, $r_{pbi}$ (Cont'd)

| Status | Number of Years |
|---|---|
| 1 | 20 |
| 1 | 24 |
| 1 | 27 |
| 0 | 29 |
| 1 | 30 |
| 1 | 31 |
| 1 | 33 |
| 1 | 34 |
| 0 | 35 |
| 1 | 36 |
| 0 | 37 |
| 0 | 38 |
| 0 | 39 |
| 0 | 40 |
| 0 | 42 |
| 0 | 44 |

(Click in this red rectangle to see SAS calculations for this example.)

This is the hypothetical Army anthropometric data for the problem described on the previous slide.

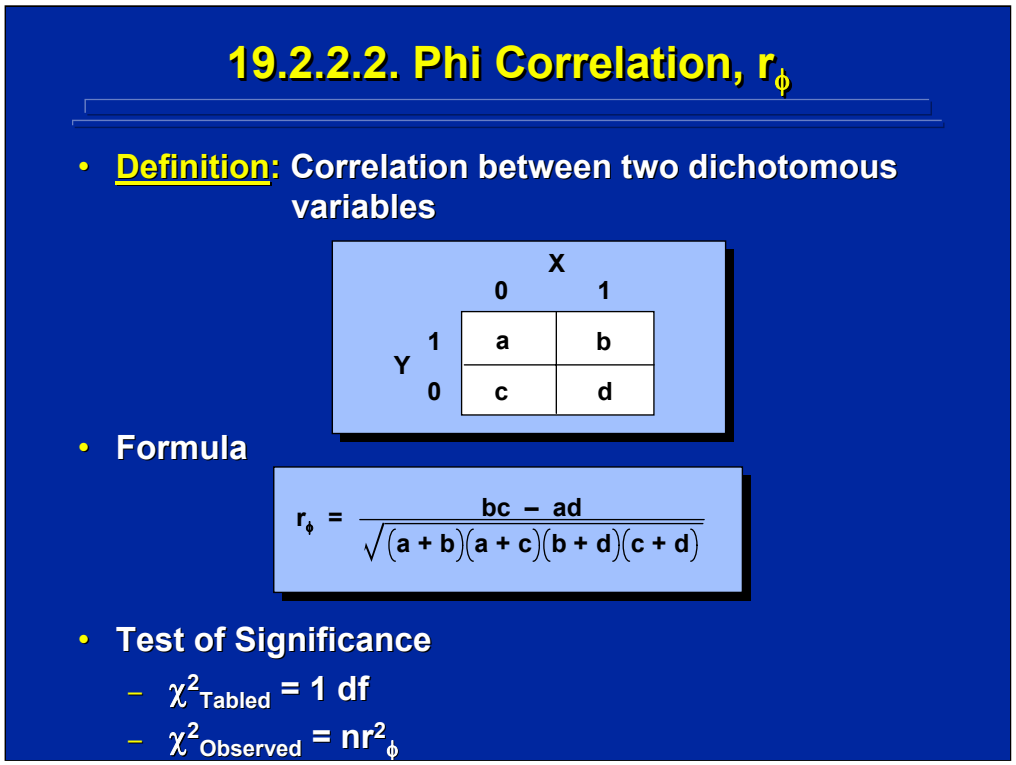
## 19.2.2.1. Point Biserial Correlation, $r_{pbi}$ (Cont'd)

$$r_{pbi} = \frac{\sum X_1 - \frac{n_1 \sum X}{n}}{\sqrt{\frac{n_1 n_0}{n}\left[\sum X^2 - \frac{(\sum X)^2}{n}\right]}}$$

$$r_{pbi} = \frac{235 - \frac{8(539)}{16}}{\sqrt{\frac{8(8)}{16}\left[18807 - \frac{(539)^2}{16}\right]}} = -0.6769$$

(Click in this red rectangle to see SAS calculations for this example.)

This slide uses the hypothetical data provided on the previous slide to conduct the point biserial correlation of 0.675.




This slide shows the formula for the phi correlation where both X and Y are dichotomous variables. A chi-squared test shown on the bottom of this slide can be used to test the significance of the phi correlation.

## 19.2.2.2. Phi Correlation, $r_\phi$ (Cont'd)

- **Example Problem: What is the correlation between the 16 soldiers' status and their gender and is this correlation significant ($p < 0.05$)?**

| Status | Gender | Status | Gender |
|---|---|---|---|
| 1 | 1 | 0 | 1 |
| 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 0 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 |
| 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 0 |

(Click in this red rectangle to see SAS calculations for this example.)

This is an example problem of using the phi correlation on the hypothetical Army anthropometric data where 0 = enlisted and 1 = officer for the status dichotomous variable, and 0 = female and 1= male for the gender dichotomous variable.

## 19.2.2.2. Phi Correlation, $r_\phi$ (Cont'd)

- **Calculation**

a = 2
b = 5
c = 5
d = 4

$$r_\phi = \frac{(5)(5) - (2)(4)}{\sqrt{(2+5)(2+5)(5+4)(5+4)}} = 0.2698$$

**Test of Significance**

$X^2_{Tabled} = 1\ df = 3.84$

$X^2_{Observed} = nr^2_\phi = 16(0.2698^2) = 1.165$

Reject $H_O$: $X^2_{Observed} > X^2_{Tabled}$

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the calculation of the phi correlation of the Army anthropometric data shown on the previous slide. As shown on the bottom portion of this slide the resulting phi correlation of 0.2698 is not significantly different than 0 based on a chi-squared test of significance ($p < 0.05$).

## 19.2.2.3. Spearman Correlation, $r_\rho$

- **Definition**: Correlation between two rank orders
  - Assumes No Tied Ranks
  - n = Number of Items Ranked
- Formula

$$r_\rho = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

where, $D = (X_{Rank} - Y_{Rank})$

- Test of Significance
  - $t_{Tabled}$ = (n–2) df

$$t_{Observed} = r_\rho \sqrt{\frac{n-2}{1 - r_\rho^{\,2}}}$$

The Spearman correlation is the correlation between two rank orders. The formula presented on this slide assumes no tied ranks and is based on the difference (D) between the ranked items (n). A t-test for evaluating the Spearman correlation is shown at the bottom of this slide.

## 19.2.2.3. Spearman Correlation, $r_\rho$ (Cont'd)

- **Tied Ranks**
  - No Adjustment Needed for Small Number of Ties
  - Just Assign Average of Tied Ranks
- **Adjusted Formula for Tied Ranks, $T_X$ and $T_Y$**

$$r_\rho = \frac{(n^3 - n) - 6\Sigma D^2 - (T_X + T_Y) / 2}{\sqrt{(n^3 - n)^2 - (T_X + T_Y)(n^3 - n) + T_X T_Y}}$$

$$T_X \text{ and } T_Y = \sum_{i=1}^{g} (t_i^3 - t_i)$$

where, $g$ = number of groupings of tied ranks.

$t_i$ = number of tied ranks in $i^{th}$ grouping.

This slide shows an adjustment for tied ranks. However, no adjustment is needed for a small number of ties. In some cases the average of the tied ranks can be used for the rank order value.

## 19.2.2.3. Spearman Correlation, $r_\rho$ (Cont'd)

- **Example Problem: The number of years of service and the remaining number of months the officers' believe they will be stationed at their post were converted into rank orders. What is the correlation between these two rank orders and is this correlation significant ($p < 0.05$)?**

(Click in this red rectangle to see SAS calculations for this example.)

This is an example problem for using a Spearman correlation to evaluate the linear correlation between two rank orders when years of Army service and months remaining at current post for Army officers are converted to rank orders.

## 19.2.2.3. Spearman Correlation, $r_\rho$ (Cont'd)

- **Problem Data**

| Years of Service (X) | Remaining Months (Y) | X Rank | Y Rank | Difference | Difference $^2$ |
|---|---|---|---|---|---|
| 8 | 1 | 8 | 1 | 7 | 49 |
| 4 | 2 | 4 | 2 | 2 | 4 |
| 5 | 3 | 5 | 3 | 2 | 4 |
| 1 | 4 | 1 | 4 | - 3 | 9 |
| 2 | 5 | 2 | 5 | - 3 | 9 |
| 3 | 6 | 3 | 6 | - 3 | 9 |
| 13 | 7 | 13 | 7 | 6 | 36 |
| 16 | 8 | 16 | 8 | 8 | 64 |
| 7 | 9 | 7 | 9 | - 2 | 4 |
| 6 | 10 | 6 | 10 | - 4 | 16 |
| 14 | 11 | 14 | 11 | 3 | 9 |
| 15 | 12 | 15 | 12 | 3 | 9 |
| 10 | 13 | 10 | 13 | - 3 | 9 |
| 9 | 14 | 9 | 14 | - 5 | 25 |
| 11 | 15 | 11 | 15 | - 4 | 16 |
| 12 | 16 | 12 | 16 | - 4 | 16 |

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the rank orders and difference scores of years of service (X) and months remaining on current duty station (Y) of the 16 soldiers used in the Spearman correlation example problem. No tied ranks appear in this example problem.

## 19.2.2.3. Spearman Correlation, $r_\rho$ (Cont'd)

- **Calculation**

$$r_\rho = 1 - \frac{6\,(288)}{16\,(16^2-1)} = 0.5765$$

- **Significance Test**

$$t_{Observed} = 0.5765\sqrt{\frac{16-2}{1-0.5765^2}} = 19.06$$

$$t_{Tabled} = (16\text{-}2) = 14\ df = 2.145$$

Reject $H_O$: $t_{Observed} > t_{Tabled}$

(Click in this red rectangle to see SAS calculations for this example.)

The Spearman correlation shown on the top portion of this slide is 0.5765 for the example data presented on the previous slide as calculated by the formula for untied ranks. The 0.5765 correlation is significant at the 0.05 level, as shown on the bottom portion of this slide. Consequently, there is a positive linear relationship between the rank order of years of service and months remaining in current duty station.

## 19.2.2.4. Partial Correlation, $r_{(1.3)(2.3)}$

- Introduction
  - Correlation Affected by Third Variable
  - Removes Variance of Third Variable from Both
  - Example: Height ($X_1$) and Weight ($X_2$) are correlated, but both are correlated with Age ($X_3$).
- Definition: $r_{(1.3)(2.3)}$ = Correlation between $X_1$ and $X_2$ with $X_3$ held constant
- Formula

$$r_{(1.3)(2.3)} = \frac{r_{12} - (r_{13})(r_{23})}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

The correlation between two other variables can be affected by the correlation of each of these variables with a third variable. A partial correlation removes the covariance of the third variable from the correlation of the first two variables. For example, height (1) and weight (2) are correlated, but they are also correlated with age (3). The correlation between height and weight when the effect of age is removed is designated as the partial correlation, $r_{(1.3)(2.3)}$. The general formula for a partial correlation is shown at the bottom of the slide. Note that three correlations are used in calculating the partial correlation coefficient.

## 19.2.2.4. Partial Correlation, $r_{(1.3)(2.3)}$ (Cont'd)

- **Test of Significance**
  - Degrees of Freedom: (n-3) not (n-2)
- **T-Test**

  $t_{Tabled}$ = (n-3) df

  $t_{Observed} = r_{(1.3)(2.3)} \sqrt{\frac{n-3}{1-r^2_{(1.3)(2.3)}}}$
- **Test Format**

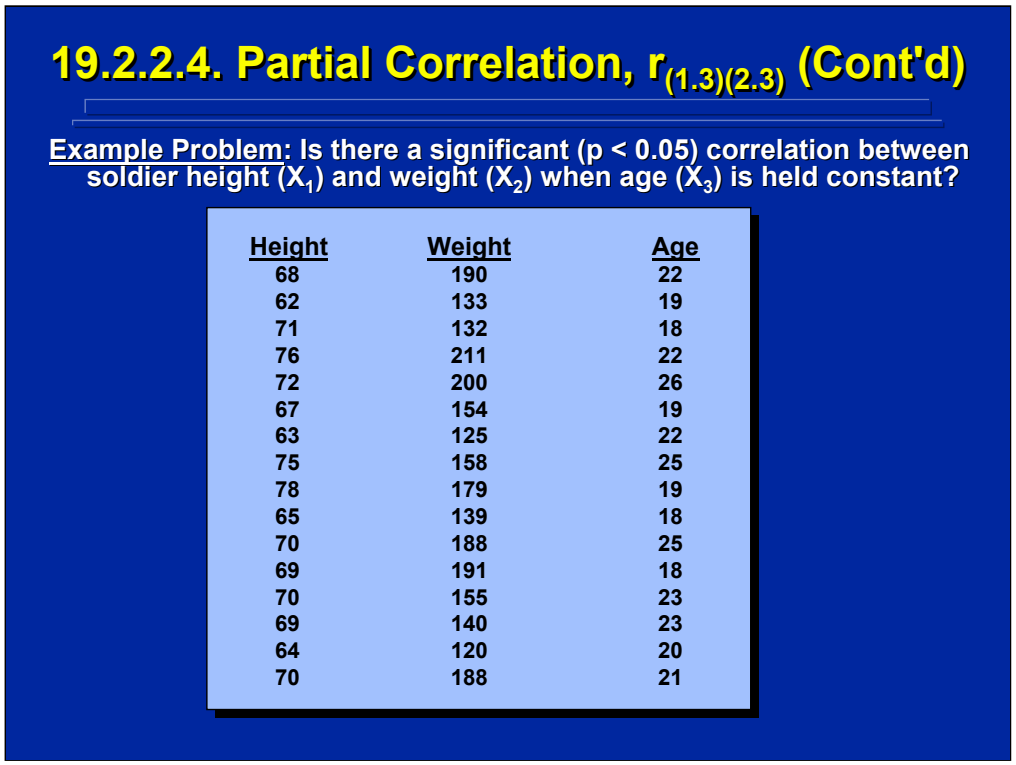
  $H_0: \rho = 0$

  $H_i: \rho \neq 0$

  $\alpha$: 0.05

  D.R.: Reject $H_0$ if $t_{Observed} > t_{Tabled}$

A t-test can be used to test the significance of partial correlations. The tabled value of the t-statistic is based on n-3 degrees of freedom because three correlations are considered in a partial correlation as shown in the formula on the previous slide. The observed value of the t-statistic is given in the middle portion of the slide, and the standard format for the t-test is shown in the bottom portion of the slide.




(Click in this red rectangle to see SAS calculations for this example.)

Hypothetical Army anthropometric data on height, weight, and age of 16 soldiers are presented on this slide to demonstrate an example calculation of a partial correlation between height (1) and weight (2) where the correlation effect of age (3) is removed.

## 19.2.2.4. Partial Correlation, $r_{(1.3)(2.3)}$ (Cont'd)

- **Partial Correlation Calculation**

$r_{12} = 0.63$
$r_{13} = 0.23$
$r_{23} = 0.35$

$$r_{(1.3)(2.3)} = \frac{0.63 - (0.23)(0.35)}{\sqrt{(1 - 0.23^2) - (1 - 0.35^2)}} = 0.59$$

- **Test of Significance**

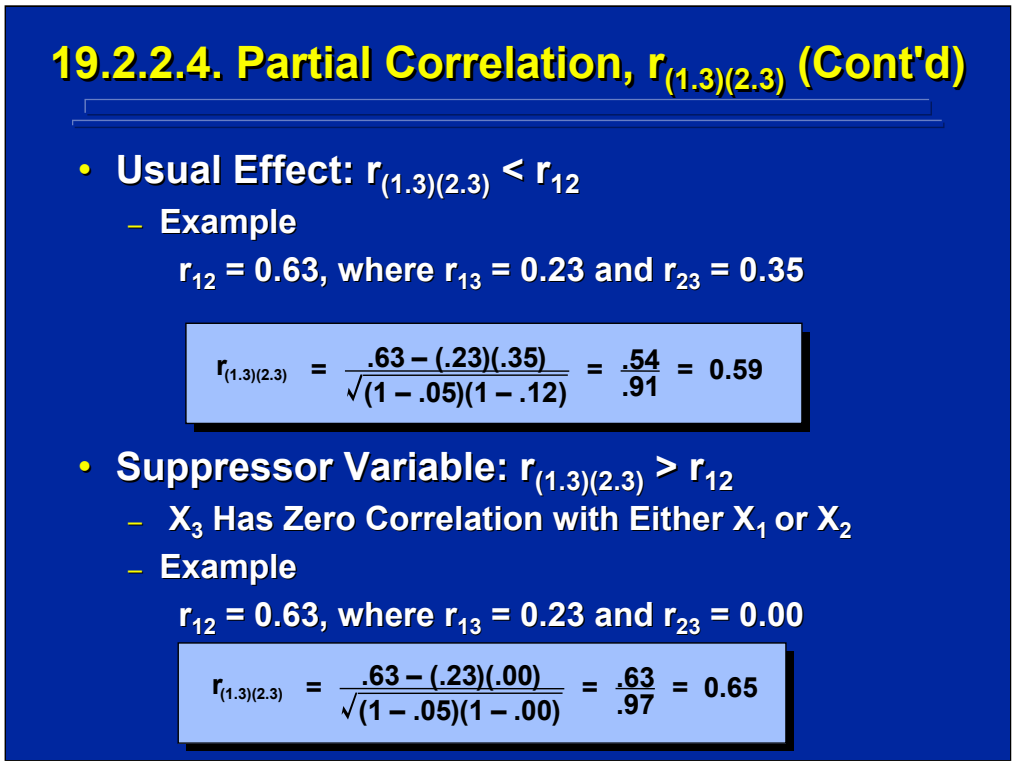
$T_{Tabled} = (n - 3\ df) = 13\ df = 2.160$

$T_{Observed} = 0.59\sqrt{13 / (1 - 0.59^2)} = 2.635$

$T_{Observed} > T_{Tabled}$ : Reject $H_O$

(Click in this red rectangle to see SAS calculations for this example.)

The calculations of the partial correlation are shown on the top of this page resulting in a partial correlation of 0.59 between soldier height and weight when the effect of age is removed. At the bottom of the slide, the results of the t-test show that this partial correlation is significant ($p < 0.05$).




(Click in this red rectangle to see SAS calculations for this example.)

The usual result is that a partial correlation between two variables is less than the simple correlation between two variables when the third effect is not considered. In this example problem, the partial correlation of height and weight of 0.59 is less than the simple correlation between height and weight of 0.63.

If one of the two variables in the partial correlation has a zero correlation with the third variable, the third variable is called a suppressor variable. For example, age becomes a suppressor variable when the simple correlation between $r_{23}$ is 0 as shown on the bottom portion of this slide. In this case, the partial correlation (0.65) is now greater than the simple correlation coefficient (0.63).

## 19.2.2.5. Semipartial Correlation, $r_{1(2.3)}$

- Introduction
  - Correlation Affected By Third Variable
  - Removes Variance of Third Variable From Only One
  - Used in Conditional Test of Significance in Multiple Regression when Predictors are Correlated
- Definition: $r_{1(2.3)}$ = Correlation between $X_1$ and $X_2$ after the variance that $X_3$ has in common with $X_2$ is removed from $X_2$
- Formula

$$r_{1(2.3)} = \frac{r_{12} - (r_{13})(r_{23})}{\sqrt{(1 - r_{23}^2)}}$$

A semipartial correlation is the linear relationship between variables when the relationship between a third variable is removed from only one of the two variables being correlated. In other words, it is the correlation between two variables where the unique contribution of the second variable is evaluated given a third variable is present. The computational formula is shown on this slide.

Semipartial correlations are used in multiple regression problems where the first variable is predicted by the second and third variables. Semipartial regressions can be used to test significance of the unique contribution of each predictor in multiple linear regression when the predictors are correlated as discussed in Topic 22.

## 19.2.2.5. Semipartial Correlation, $r_{1(2.3)}$ (Cont'd)

- **Example Problem**: Soldier Height ($X_1$) is predicted by both Weight ($X_2$) and Age ($X_3$). What is the unique contribution of Weight given Age is included in the data analyzed on the 16 soldiers?

$$r_{1(2.3)} = \frac{r_{12} - (r_{13})(r_{23})}{\sqrt{(1 - r_{23}^2)}}$$

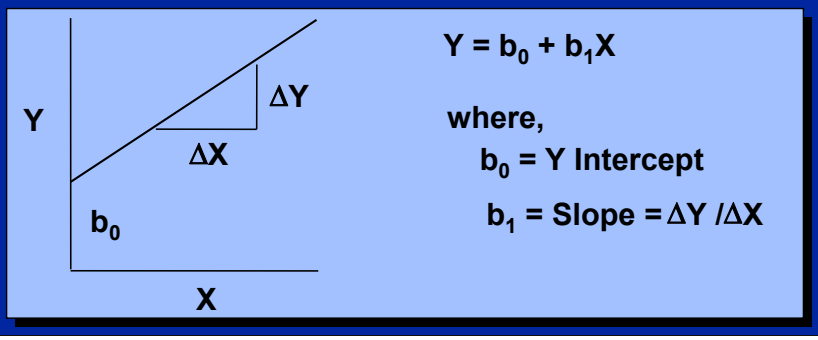
$$r_{1(2.3)} = \frac{0.63 - (0.23)(0.35)}{\sqrt{(1 - 0.35^2)}} = \frac{.54}{.94} = 0.56$$

(Click in this red rectangle to see SAS calculations for this example.)

This slide provides an example of a semipartial correlation using the previous anthropometric data of height, weight, and age of 16 soldiers. Specifically, the correlation between height and weight is 0.63 when age is not considered in the prediction of height. But, the correlation between height and the unique contribution of weight given age is also included in predicting height is only 0.56. Consequently, some of the correlation between height and weight is due to the covariance of age with height.

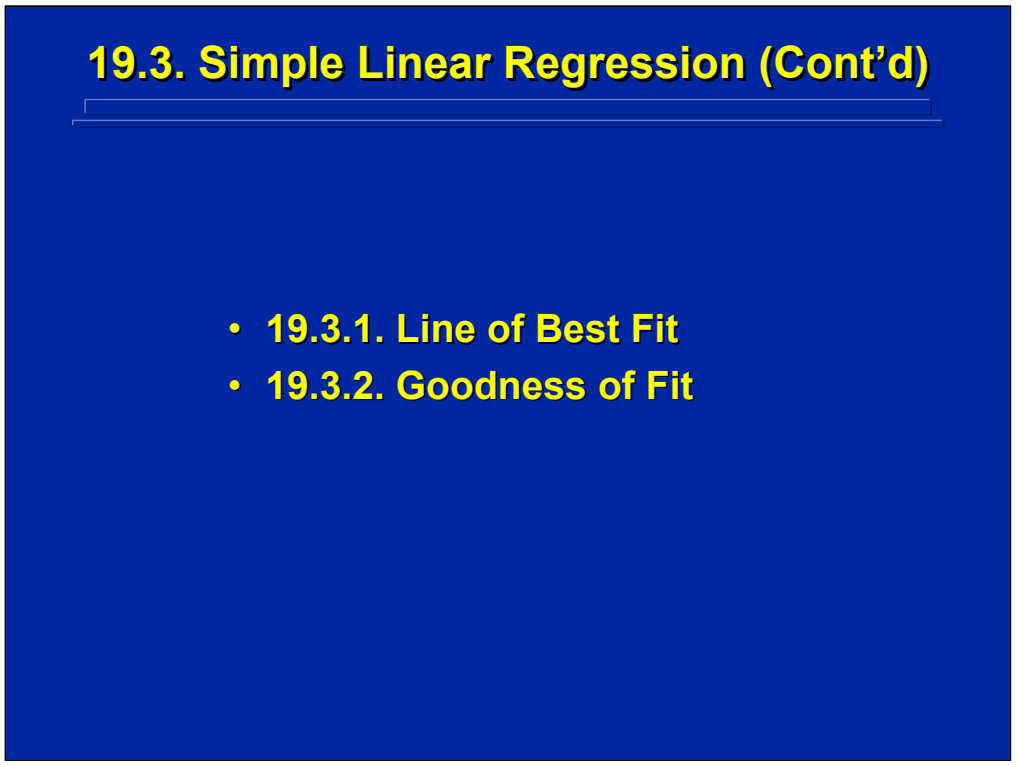
## 19.3. Simple Linear Regression

- **Definition: Regression is the method by which a value of one variable, Y, can be predicted by knowing the value of another variable, X.**
  - **Assumes a Linear Relationship**
  - **Correlation Exists Between X and Y**
- **General Linear Model**

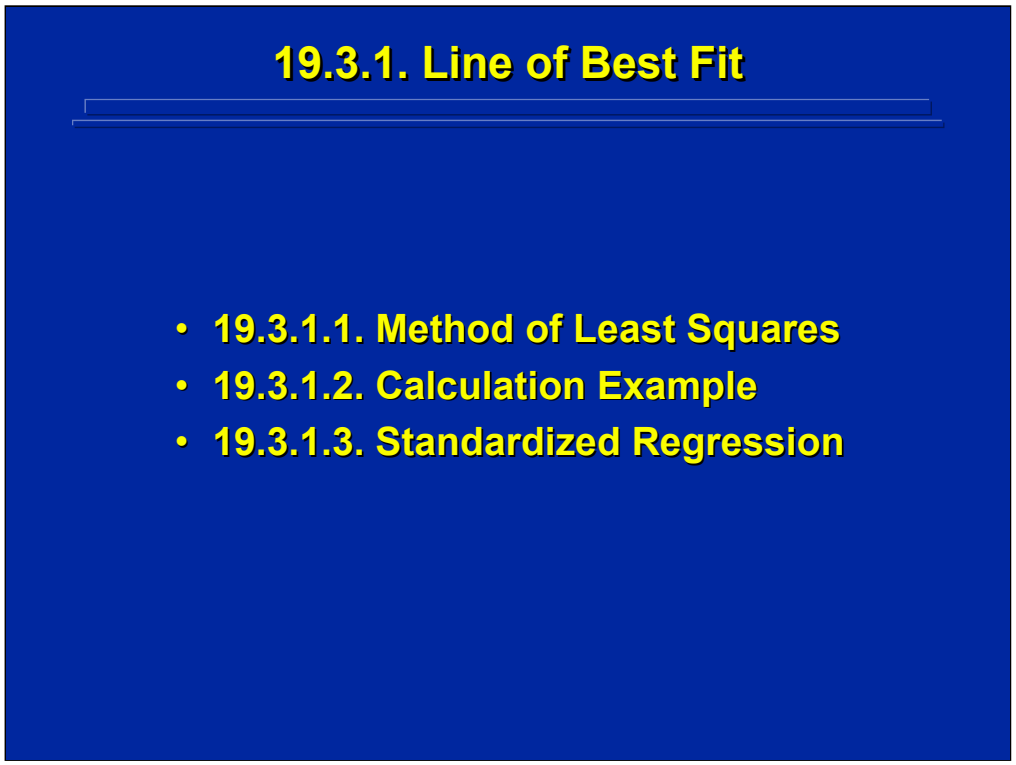



Regression is used to predict one variable (Y) as a function of another variable (X). Simple regression assumes a linear relationship between the two correlated variables. If the two variables were not correlated at all, the linear model would be horizontal. If the two variables were perfectly correlated, then all the XY data points would fall directly on the line describing the linear relationship. In reality, deviation from the predicted linear model can be used to assess the goodness of fit of the simple linear regression model.

As shown on this slide the simple linear regression of Y as a function of X can be written as $Y = b_0 + b_1X$ with two parameters $b_0$ and $b_1$ where $b_0$ is the Y intercept and $b_1$ is the slope of the line. Sample data of Y and X values are collected to solve the two regression weights in the simple linear regression model.




Two general categories of calculations are conducted in simple regression. First, the line of best fit is determined based on the sample data. This involves solving the parameters $b_0$ and $b_1$ in the linear model for simple regression. Second, the goodness of fit of the simple linear regression is evaluated in order to assess the adequacy of the linear model. Techniques for conducting each of these two categories of computation are described separately in this subsection.




Calculating the line of best fit in simple linear regression uses the least squares criterion for solving the two parameters $b_0$ and $b_1$ in the regression line. Both raw score and standardized, Z, score solutions are discussed.

## 19.3.1.1. Method of Least Squares

- **Observed vs. Predicted Scores**
  - Y = Observed Score
  - Y' = Predicted Score
  - where, $Y' = b_0 + b_1X$
- **Error in Prediction**
  - $Y - Y' = Y - (b_0 + b_1X)$
- **Least Squares Criterion**
  - Determine Value of "$b_0$" and "$b_1$" such that the sum of the squared difference between (Y – Y') is a minimum.

$$\sum (Y - Y')^2 = \sum [Y - (b_0 + b_1X)]^2 = \text{minimum}$$

In simple regression the predicted score, Y', is determined by the linear equation, $Y' = b_0 + b_1X$. The difference between the observed score, Y, and the predicted score, Y', is a measure of the error in regression. The least squares criterion keeps the sum of the squared differences between Y-Y' at a minimum in solving the two parameters $b_0$ and $b_1$ in simple linear regression.

## 19.3.1.1. Method of Least Squares (Cont'd)

- **Least Squares Solution: Take partial derivatives with respect to "$b_0$" and "$b_1$"; set the derivatives equal to zero; and solve for "$b_0$" and "$b_1$".**
- **Simultaneous Equations**

$$\left\{ \begin{array}{l} \Sigma Y = nb_0 + b_1 \Sigma X \\ \Sigma XY = b_0 \Sigma X + b_1 \Sigma X^2 \end{array} \right\}$$

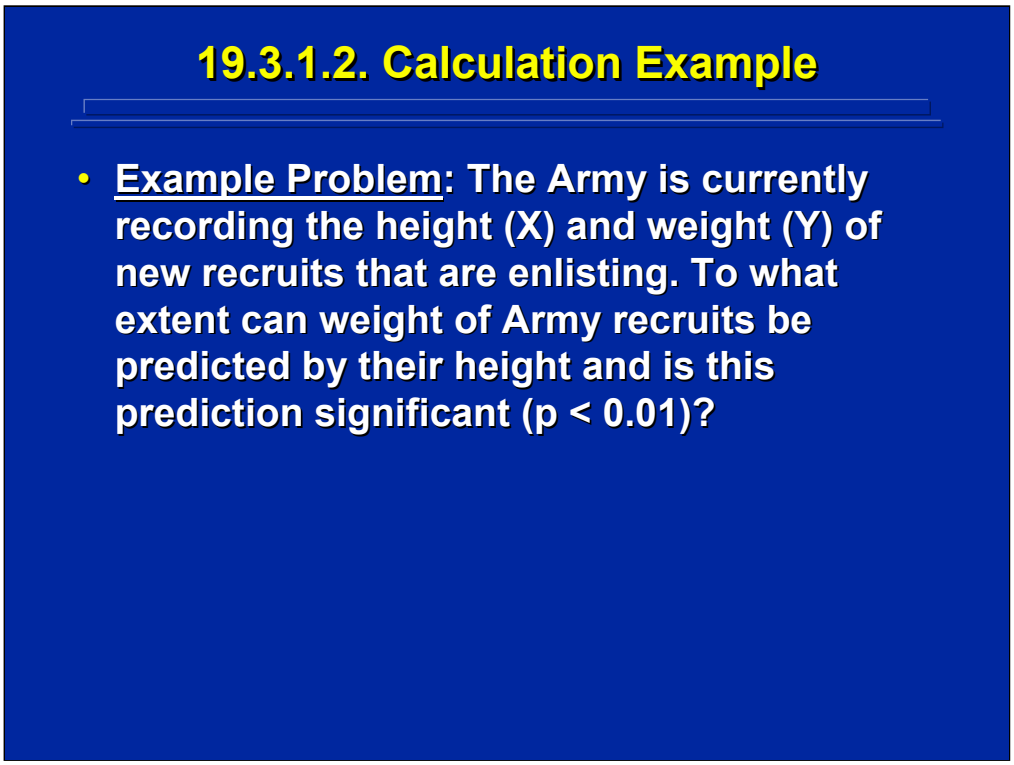
- **Value of "$b_0$"**

$$b_0 = \bar{Y} - b_1 \bar{X}$$

- **Value of "$b_1$"**

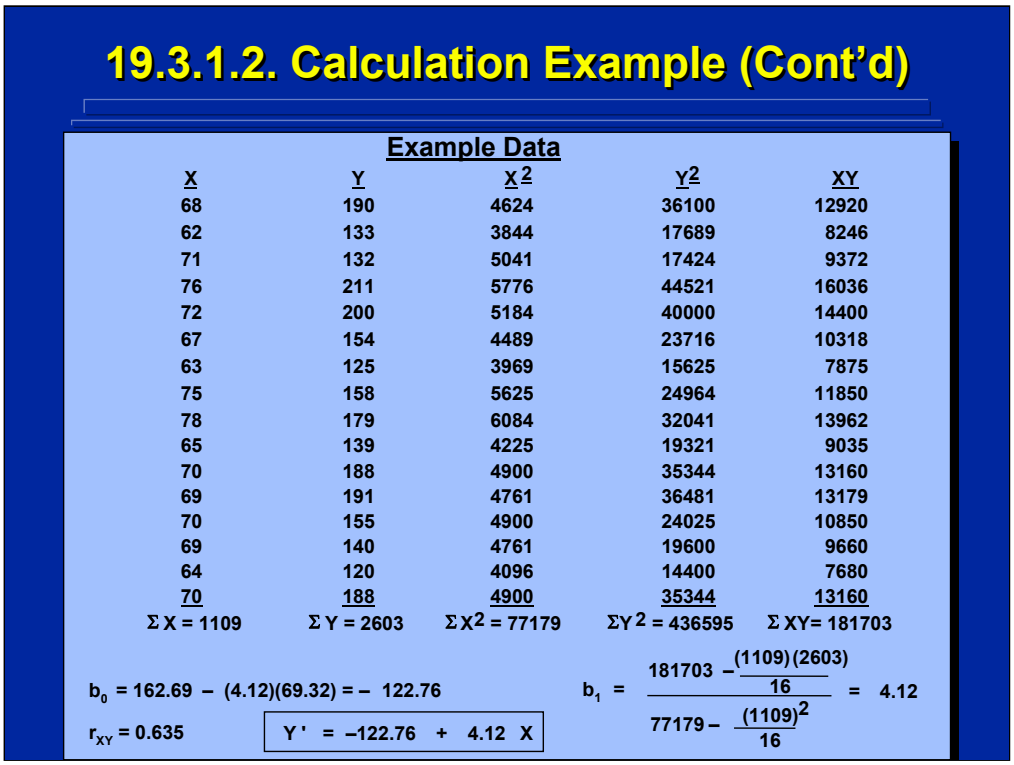
$$b_1 = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}$$

The least square criterion requires that partial derivatives of the regression line be calculated with respect to $b_0$ and $b_1$ resulting in the two simultaneous equations shown in parenthesis in this slide. Solving these two simultaneous equations for $b_0$ and $b_1$ provides the least squares criterion solution for the two parameters in simple linear regression. The solution of these two simultaneous equations provides the formulae for computing the values for $b_0$ and $b_1$ are shown on the bottom half of this slide.



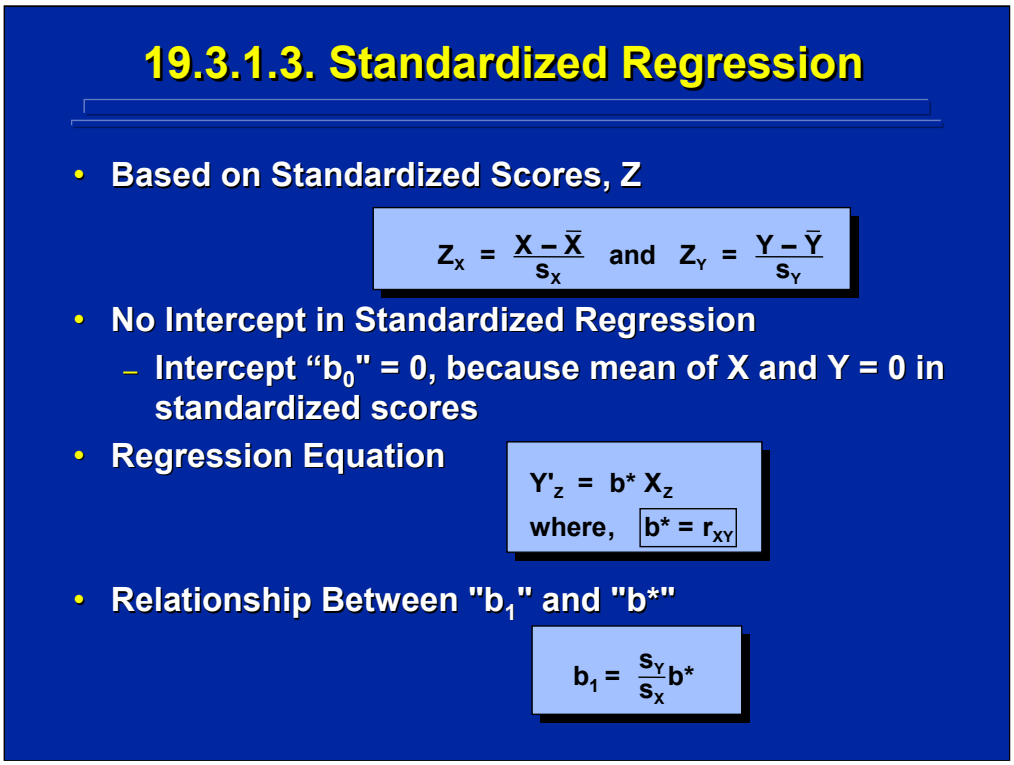

(Click in this red rectangle to see SAS calculations for this example.)

This slide states a simple linear regression problem in which the Army wants to predict soldier weight (Y) as a function of soldier height (X) and determine if this prediction is statistically significant at the 0.01 level of significance. The Slater and Williges (2006) appendix provides the SAS program solution to this simple regression example.

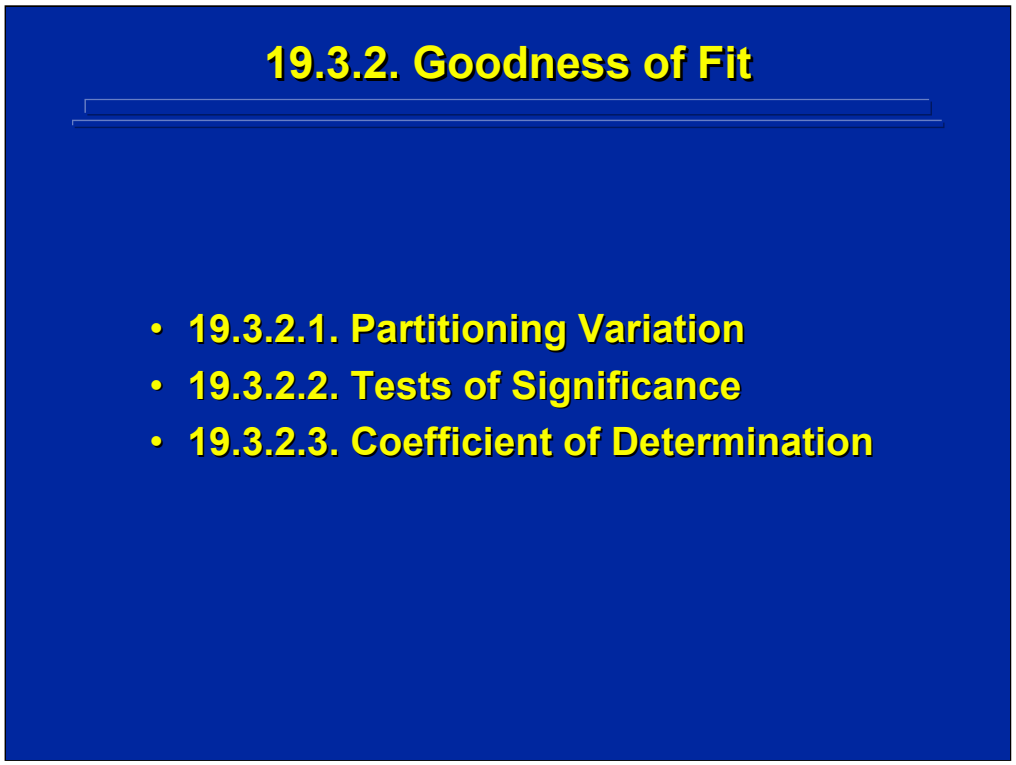



(Click in this red rectangle to see SAS calculations for this example.)

This slide presents the hypothetical data from 16 soldiers of their height (X) in inches and their weight (Y) in pounds. The resulting simple linear regression for predicting weight, Y', as a function of height (X) is shown in the box at the bottom of this slide. The values for $b_0$ and $b_1$ were calculated using the formulae for a least squares solution presented in a previous slide. These calculations yield an intercept of -122.76 and a slope of 4.12 for the simple linear regression line. The correlation between the 16 X and Y scores in this sample is 0.635.

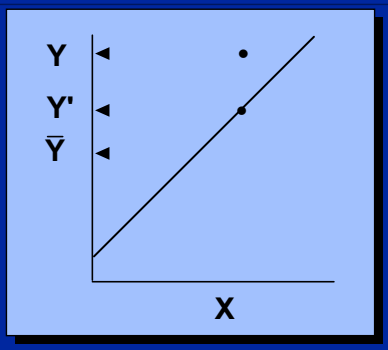



An alternative to the raw score regression line is a standardized regression stated in terms of Z scores. If Z scores are used, the intercept is forced to zero and the $b_0$ parameter drops out of the regression equation. As shown on this slide the value for $b_1$ is stated as $b^*$ in standardized regression and is equal to the correlation between X and Y. For the example problem, the standardized regression equation is simply $Y' = 0.636X_Z$ based on the correlation provided on the previous slide.




Goodness of fit of the regression equation is based on the partitioning of variation between the observed value of Y and the predicted value Y'. Both a test of significance and a coefficient of determination can be used to assess the goodness of fit of a regression equation as discussed in this subsection.

## 19.3.2.1. Partitioning Variation




- Additive Components of Y Score Deviation, $(Y - \bar{Y})$
  - Deviation Due to Regression, $(Y' - \bar{Y})$
  - Deviation Due to Error of Estimation, $(Y - Y')$
- Variation of Y Scores

$$\sum (Y - \bar{Y})^2 = \sum (Y' - \bar{Y})^2 + \sum (Y - Y')^2$$

The deviation of a score Y from its mean can be divided into two additive parts when considering the predicted Y' value in regression. First, there is the deviation due to regression which is the difference between Y' and the mean. And, second, there is the deviation due to error in estimation which is the difference between Y and Y'. As shown on the bottom of this slide, the sum of squares (SS) of deviations can also be divided into these two additive parts. Namely, total SS can be divided into Regression SS and Error SS which can be used in an ANOVA to test the goodness of fit of regression.

## 19.3.2.2. Tests of Significance

- **ANOVA on Simple Linear Regression Model**
  - Tests only $b_1$ when Corrected for Mean
  - Test of Linear Association Between X and Y
  - $F_{(1, n-2)} = MS_{Regression} / MS_{Error} = t_r^2$
  - $Ms_{Regession}$ = Regression Model Variance
  - $MS_{Error}$ = Deviation from Model Variance ($s^2$)
- **Format for Simple Regression ANOVA**
  - $H_0$: $\beta = 0$
  - $H_i$: $\beta \neq 0$
  - $\alpha$: 0.05, 0.01, or 0.001
  - D.R.: I reject $H_0$ if $F_{Observed} > F_{Tabled}$

An ANOVA on regression tests the goodness of fit of the linear association between X and Y. In simple regression only the slope of the regression line, $b_1$, is tested when the data are corrected for the sample mean in the ANOVA. The resulting F-ratio is simply $MS_{Regression}$ divided by $MS_{Error}$ with 1 and n-2 degrees of freedom. Since the F-test on $b_1$ has only 1 degree of freedom, it is equivalent to a $t^2$ test of regression. The standard format for testing the significance of regression is given in the bottom part of this slide.

## 19.3.2.2. Tests of Significance (Cont'd)

### ANOVA Summary Table (Corrected for Mean)

| Source | df | Sum of Squares | MS | F |
|---|---|---|---|---|
| Regression (R) | 1 | $\sum (Y' - \bar{Y})^2 = b_1 \sum xy$ | $MS_R$ | $MS_R/MS_E$ |
| Error (E) | n-2 | $\sum (Y - Y')^2 = \sum y^2 - b_1 \sum xy$ | $MS_E$ | |
| Total | n-1 | $\sum (Y - \bar{Y})^2 = \sum y^2$ | | |

Where Raw Score Equivalents,

$$\Sigma_{xy} = \Sigma_{XY} - \frac{\Sigma X \Sigma Y}{n}$$

$$\Sigma_{x^2} = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

$$\Sigma_{y^2} = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$$

This slide shows the ANOVA summary table for a test of significance of simple regression. The SS of the deviation components of the raw scores are also listed in terms of $b_1$ and sums of deviation scores that are defined at the bottom of the ANOVA Summary Table as described by Myers (1990, pp 22-29).

## 19.3.2.2. Tests of Significance (Cont'd)

**ANOVA Summary Table (Corrected for Mean)**

| Source | df | Sum of Squares | MS | F |
|---|---|---|---|---|
| Regression (R) | 1 | 5281.85 | 5281.85 | 9.43* |
| Error (E) | 14 | 7837.59 | 559.83 | |
| Total | 15 | 13119.44 | | |

*p < 0.01

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the complete ANOVA Summary Table that is calculated by the formulae shown on the previous slide for the example simple regression problem data. Regression is statistically significant ($p < 0.01$) which means that the regression line predicting soldier weight from soldier height has a significant linear relationship.

## 19.3.2.2. Tests of Significance (Cont'd)

- **t-Test on Individual Beta Weights**
  - t-Test on $b_i$, where $t_{Observed} = \frac{b_i}{s_{bi}}$
  - $S_{bi}$ = Standard Error of a Particular Beta Weight, $b_i$
  - t-Test on $b_1$ = t-Test on $r_{XY}$ in Simple Regression
- **Format for t-Test using $r_{XY}$ in Simple Regression**
  - $H_0$: $\beta_1 = 0$
  - $H_i$: $\beta_1 \neq 0$
  - $\alpha$: 0.05, 0.01, or 0.001
  - D.R.: I reject $H_0$ if $t_{Observed} > t_{Tabled}$

$$t_{Observed} = |r|\sqrt{\frac{n-2}{1-r^2}} = .635\sqrt{\frac{16-2}{1-(.635)^2}} = 3.072$$

$t_{Tabled}$ = (n-2) df = 2.145 (14 df)

(Click in this red rectangle to see SAS calculations for this example.)

A t-test could be used as an alternative to the Regression ANOVA since the test of a beta weight has only 1 degree of freedom. The $t_{Observed}$ value is simply the beta weight divided by its standard error. Alternatively, Hays (1994, p. 648) demonstrates that the test on $b_1$ is equivalent to testing the significance of the correlation between X and Y in simple regression.

The format for t-test on $b_1$ using $r_{XY}$ in simple regression is shown on the bottom portion of this slide. Note that the t-test on the data from the example problem is significant ($p < 0.01$) just as shown in the ANOVA of regression on the previous slide.

## 19.3.2.2. Tests of Significance (Cont'd)

- $t_{Observed}$ for $b_0$ Based on Standard Error Estimate
  - Estimate of Error Variance: $s_{b0}^2 = \sum y^2(1/n + \bar{X}^2/\sum x^2)$
  - Example: $b_0 = -122.76$
    - $s_{b0}^2 = 8670.89$ and $s_{b0} = 93.12$
    - $t_{Observed} = b_0 / s_{b0} = -1.32$ ($p>0.01$)
- $t_{Observed}$ for $b_1$ Based on Standard Error Estimate
  - Estimate of Error Variance: $s_{b1}^2 = (\sum y^2) / (\sum x^2)$
  - Example: $b_1 = 4.12$
    - $s_{b1}^2 = 1.80$ and $s_{b1} = 1.34$
    - $t_{Observed} = b_1 / s_{b1} = 3.07$ ($p < 0.01$)
  - $t_{Observed}$ for $r_{XY}$ = $t_{Observed}$ for $b_1 = 3.07$
  - $F_{Regression} = (t_{Observed})^2 = 9.43$

(Click in this red rectangle to see SAS calculations for this example.)

Myers (1990, pp 14-19) provides the formulae for the error variances of $b_0$ and $b_1$ shown on this slide. The standard error of these variances can be used in a t-test on the significance of both the intercept and the slope of simple regression, respectively. Both the intercept and slope are significant ($p < 0.01$) in the example problem.

Note that the bottom portion of this slide shows that the $t_{Observed}$ on $b_1$ given on this slide (i.e., 3.07) is the same value as the t-test on the $r_{XY}$ shown on the previous slide. In addition, the square of this $t_{Observed}$ value is the same as the F-ratio in the Regression ANOVA Summary Table (i.e., 9.43).

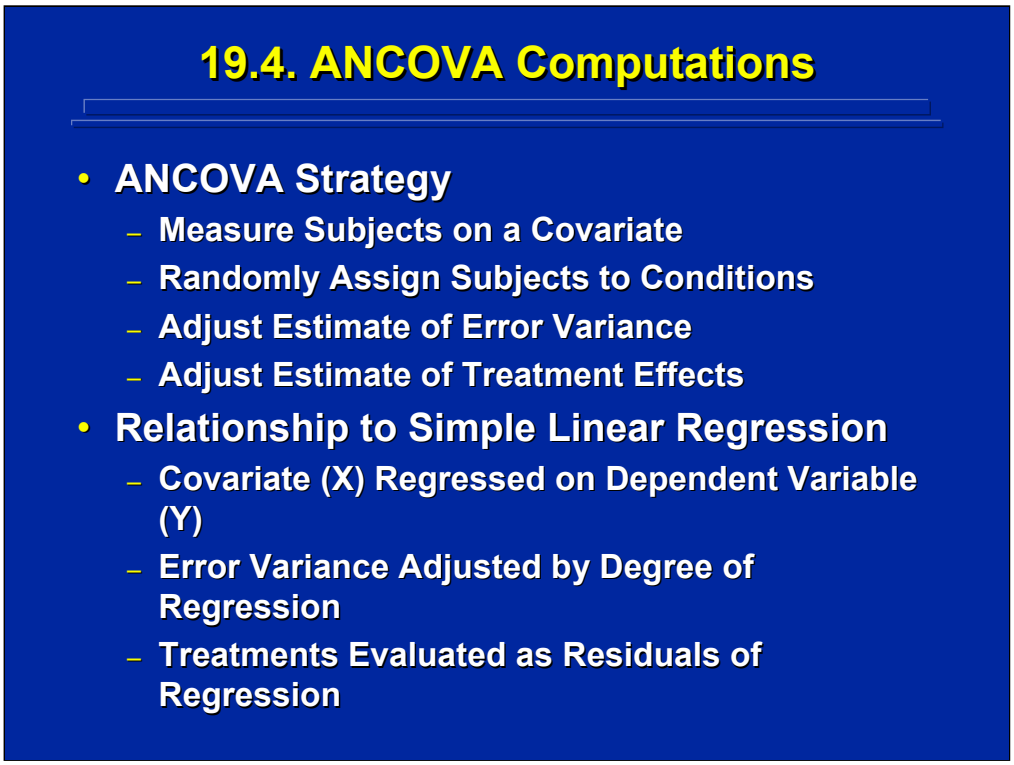
## 19.3.2.3. Coefficient of Determination

- Multiple Correlation Coefficient: $R = r_{YY'}$
- Coefficient of Determination = $R^2$
  - Definition: Percent of total variation predicted by regression model.
- Calculation of $R^2$
  - Square of Multiple Correlation Coefficient, R
    $R^2 = (r_{YY'})^2$
    $r_{YY'} = 0.6345$
    $R^2 = (0.6345)^2 = 0.4026$
  - ANOVA on Regression
    $R^2 = SS_{Regression} / SS_{Total}$
    $R^2 = 5281.85/13119.44 = 0.4026$
  - Example: 40.26% of Total Variation Predicted by Regression Model

(Click in this red rectangle to see SAS calculations for this example.)

Another way to determine the goodness of fit of a regression equation is to determine the percent of total variation predicted by the regression model. The multiple correlation coefficient, R, is the correlation between Y and Y', the predicted value of Y (i.e., $r_{YY'}$). The square of the multiple correlation coefficient is called the Coefficient of Determination ($R^2$). Myers (1990, p. 37) defines $R^2$ as the $SS_{Regression}/SS_{Total}$ or the percent of variation predicted by the regression model.

For the example problem data, $r_{YY'} = 0.6345$. The square of 0.6345 is 0.4026 which is equal to $R^2$ or $SS_{Regression}/SS_{Total}$. Consequently, the simple regression of the sample problem accounts for 40.26% of the total variation in the regression model.



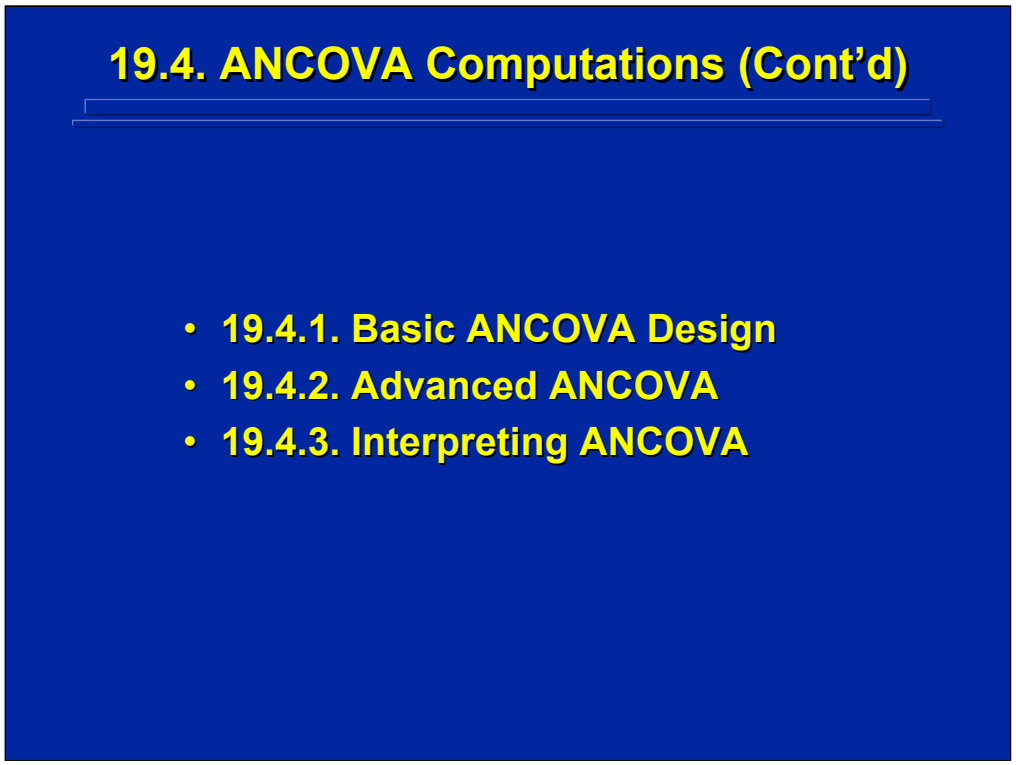

This slide summarizes the overall strategy for using ANCOVA as a means of removing the effect of a covariate from the treatment effects in a between-subjects design. This strategy results in an adjusted estimate of error variance used to test adjusted treatment means. Essentially, the covariate (X) or measure of subject differences that is correlated with the dependent variable (Y) is evaluated through simple linear regression, and the variance due to regression is removed from the ANCOVA. The subsequent analyses of treatment effects are evaluated as residuals of regression using an error term adjusted for regression.

## 19.4. ANCOVA Computations (Cont'd)

- **Covariate, X**
  - Characteristic of Subjects Correlated with Dependent Variable
  - Measure of Individual Differences
  - Usually a Classification Variable (e.g., Age, Experience, Aptitude, Attitude, etc.)
- **Assumptions of ANCOVA**
  - Random Distribution of Level of Covariate Across Treatments
  - Linear Regression Between Covariate and Dependent Variable
  - Homogeneous Group Regression Coefficients

Choice of the appropriate covariate (X) is critical to ANCOVA, and it must represent a characteristic of subjects that is significantly correlated with the dependent variable in order to remove a significant source of variance from the error term of the between-subjects design. A quantitative classification variable such as age, aptitude, or experience level of the subject is usually chosen as a covariate when it is known to be correlated with the dependent variable.

The three primary assumptions implicitly made in an ANCOVA are listed at the bottom of this slide. First, the distribution of the covariate effects across treatments is assumed to be random. Second, the significance of the simple linear regression between the dependent variable (Y) and the covariate (X) is assumed to be significant. And, third, the regression coefficients of each treatment level are assumed to be homogeneous across treatment levels.




Calculations of a basic ANCOVA design are described in this final subsection and extensions to advanced ANCOVA are discussed. In addition, a caution on interpreting ANCOVA results in terms of adjusted means is noted.

## 19.4.1. Basic ANCOVA Design

- **Example Problem: An experiment was conducted to study the effects of three different weight training methods used during basic training. One group of eight soldiers used basic weight training ($A_1$), another group of eight soldiers received weight training and aerobic exercise ($A_2$), and a third group of eight soldiers received weight training and diet control ($A_3$). The maximum lifting weight (MLW) of the 24 soldiers was measured after two months of training on one of the three methods. A covariate, the weight of each subject was measured before measurement of MLW. Were the three different weight training methods significantly different ($p < 0.05$) in terms of MLW?**

(Click in this red rectangle to see SAS calculations for this example.)

This slide describes a simple one-way, between-subjects design in which the effects of three different weight training procedures were evaluated on eight different soldiers in each training condition. The weight of each of the 24 soldiers was used as the covariate in the experiment. The appendix by Slater and Williges (2006) provides the SAS solutions for this ANCOVA example.

## 19.4.1. Basic ANCOVA Design (Cont'd)

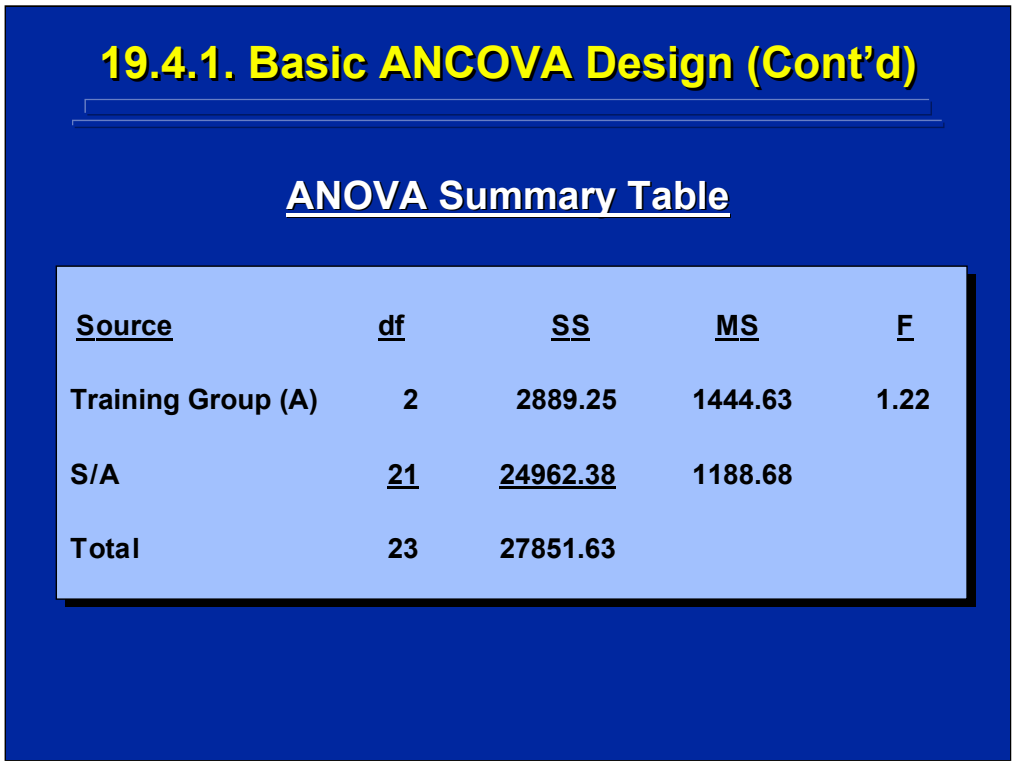
**One-Way Completely Randomized Design**
**(X = Covariate and Y = Dependent Variable)**

| | $a_1$ | | $a_2$ | | $a_3$ | |
|---|---|---|---|---|---|---|
| | X | Y | X | Y | X | Y |
| | 183 | 240 | 200 | 360 | 182 | 275 |
| | 168 | 264 | 207 | 295 | 194 | 307 |
| | 220 | 300 | 172 | 260 | 179 | 240 |
| | 200 | 342 | 188 | 305 | 213 | 333 |
| | 192 | 249 | 201 | 340 | 194 | 248 |
| | 178 | 277 | 177 | 285 | 185 | 232 |
| | 185 | 285 | 171 | 290 | 183 | 267 |
| | 190 | 263 | 167 | 255 | 193 | 289 |
| Means | 189.50 | 277.50 | 185.38 | 298.75 | 190.38 | 273.88 |
| $\Sigma X$ or $\Sigma Y$ | 1516 | 2220 | 1483 | 2390 | 1523 | 2191 |
| $\Sigma X^2$ or $\Sigma Y^2$ | 288986 | 623384 | 276637 | 723200 | 290769 | 608501 |
| $\Sigma(X)(Y)$ | 422481 | | 446085 | | 419167 | |

$T_X$ = 1516 + 1483 + 1523 = 4522
$T_Y$ = 2220 + 2390 + 2191 = 6801

(Click in this red rectangle to see SAS calculations for this example.)

This slide provides a hypothetical data set for the simple, one-way, completely randomized design described on the previous slide. The three levels of A are the three methods of weight training evaluated in the experiment with a sample size (n) equal to 8 resulting in a total of 24 different subjects in the between-subjects design. The X variable is a measure of the covariate, soldier's weight. The Y values are the maximum lifting weight (MLW) dependent measures after two months of training.




(Click in this red rectangle to see SAS calculations for this example.)

This is the standard ANOVA Summary Table for the previous slide of the MLW data (Y). Note that the main effect of Training Group (A) is not significant ($p>.05$). The effect of the covariate X (i.e., weight of the soldier) is not removed in this analysis and can contribute to the error term in this between-subjects design.

## 19.4.1. Basic ANCOVA Design (Cont'd)

### ANOVA of Simple Linear Regression (Y Predicted By X)

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Regression (R) | 1 | 9111.02 | 9111.02 | 10.70 ** |
| Error (E) | 22 | 18740.61 | 851.85 | |
| Total | 23 | 27851.63 | | |

**p < 0.01

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the result of an ANOVA conducted on the simple linear regression predicting the Y dependent variable, MLW, as a function of the X covariate, weight of soldier. The regression model is significant ($p < 0.01$) which means that X and Y are correlated and a significant amount of covariance can be removed from the one-way, between-subjects design by removing the regression effect.

## 19.4.1. Basic ANCOVA Design (Cont'd)

- **ANCOVA Calculations**
  - Calculate and Discard SS Regression
  - Calculate Adjusted SS from Regression Error
- **Sum of Squares and Sum of Products**

| Source | SS Covariate (X) | SS Depend. Var. (Y) | SP of (X)(Y) |
|---|---|---|---|
| A | $SS_{A(X)}$ = 114.08 | $SS_{A(Y)}$ = 2889.25 | $SP_{A(XY)}$ = - 573.88 |
| S/A | $SS_{S/A(X)}$ = 4257.75 | $SS_{S/A(Y)}$ = 24962.38 | $SP_{S/A(XY)}$ = 6885.13 |
| Total (T) | $SS_{T(X)}$ = 4371.83 | $SS_{T(Y)}$ = 27851.63 | $SP_{T(XY)}$ = 6311.25 |

(Click in this red rectangle to see SAS calculations for this example.)

The ANCOVA calculations disregard, or remove, the SS due to regression and restrict the analysis to the SS in the deviation from regression (i.e., Error shown on the previous slide). This regression error is used to calculate an adjusted treatment effect and error term in the one-way ANOVA.

These adjustments are made by considering the SS of the X covariate, the Y dependent variable, and the XY sum of products (SP) of X and Y deviations from their respective means as shown on the bottom portion of this slide. The SS of X and Y must be positive by definition but the SP value can be negative as shown on this slide because it is a cross product not a squared value. Note the SS for Y is exactly the same as the values shown in the ANOVA Summary Table on Y in a previous slide.

## 19.4.1. Basic ANCOVA Design (Cont'd)

- Adjusted $SS_A$ [$SS_{A(adj.)}$]
  - Formula

$$SS_{A(adj.)} = SS_{A(Y)} - \frac{(SP_{T(XY)})^2}{SS_{T(X)}} + \frac{(SP_{S/A(XY)})^2}{SS_{S/A(X)}}$$

- Adjusted $SS_{S/A}$ [$SS_{S/A(adj.)}$]
  - Formula

$$SS_{S/A(adj.)} = SS_{S/A(Y)} - \frac{(SP_{S/A(XY)})^2}{SS_{S/A(X)}}$$

- Adjusted $SS_{Total}$ [$SS_{Total(adj.)}$]
  - Formula

$$SS_{Total\ (adj.)} = SS_{T(Y)} - \frac{(SP_{T(XY)})^2}{SS_{T(X)}} = SS_{Regression\ Error\ (E)}$$

(Click in this red rectangle to see SAS calculations for this example.)

Myers (1979, pp 407-418) shows how the SS values on the previous page can be used to adjust $SS_A$ and $SS_{S/A}$ in the ANCOVA according to the formulae shown on this slide. The total of both of these adjusted SS values equals the $SS_{Regression\ Error}$ in the ANOVA on simple linear regression. Consequently, the ANCOVA on adjusted SS is restricted to the regression error SS.

## 19.4.1. Basic ANCOVA Design (Cont'd)

### ANCOVA Summary Table

| Source | df | $SS_{(adj.)}$ | $MS_{(adj.)}$ | F |
|---|---|---|---|---|
| Training Group (A) | 2 | 4912.03 | 2456.02 | 3.55* |
| S/A | 20 | 13828.58 | 691.43 | |
| Total | 22 | 18740.61 | | |

*p < 0.05

(Click in this red rectangle to see SAS calculations for this example.)

The ANCOVA Summary Table for the soldier weight training example data is shown on this slide. Now Factor A (Training Group) is significant ($p < 0.05$). So, adjusting for the covariate X in the ANCOVA provided a more sensitive test of Factor A than the ANOVA shown on a previous slide that failed to show significance.

Note that the total adjusted SS for Factor A and S/A is the same as the SS for Error in the ANOVA simple regression (i.e., 18740.61). Also notice that the degrees of freedom for S/A in the ANCOVA on this slide is 20; whereas, the degrees of freedom for S/A in the ANOVA on a previous slide is 21. The 1 df difference in the ANOVA and the ANCOVA is the 1 df removed from the error term by the simple regression of the covariate considered in the ANCOVA.

## 19.4.1. Basic ANCOVA Design (Cont'd)

### Complete Breakdown of Sum of Squares

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Regression (R) | 1 | 9111.02 | 9111.02 | 10.70 ** |
| Error (E) | (22) | (18740.61) | 851.85 | |
| A | 2 | 4912.03 | 2456.02 | 3.55* |
| S/A | 20 | 13828.58 | 691.43 | |
| Total | 23 | 27851.63 | | |

*p < 0.05 and **p < 0.01

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the complete ANCOVA Summary Table that occurs in the ANCOVA, showing both the significant ($p < 0.01$) regression effect and the breakdown of regression error used in the ANCOVA to determine the significant ($p < 0.05$) Factor A effect based on adjusted means. Post hoc testing is required to isolate the main effect of Factor A. Remember the means for Factor A must be interpreted as adjusted for the covariate effect not as the original means of Factor A evaluated in the ANOVA.

## 19.4.1. Basic ANCOVA Design (Cont'd)

- **Adjust Treatment Means for Chance Differences on Covariate, X**
- **Formula**

$$\bar{Y}'_{A_i} = \bar{Y}_{A_i} - b_{S/A}(\bar{X}_{A_i} - \bar{X}_T)$$

where,

$\bar{Y}'_{A_i}$ = the adjusted treatment mean for level $a_i$

$\bar{Y}_{A_i}$ = the unadjusted treatment mean for level $a_i$

$\bar{X}_{A_i}$ = the group mean on the covariate for level $a_i$

$\bar{X}_T$ = the grand mean on the covariate

$$b_{S/A} = \frac{SP_{S/A}}{SS_{S/A(X)}}$$

- **Example**

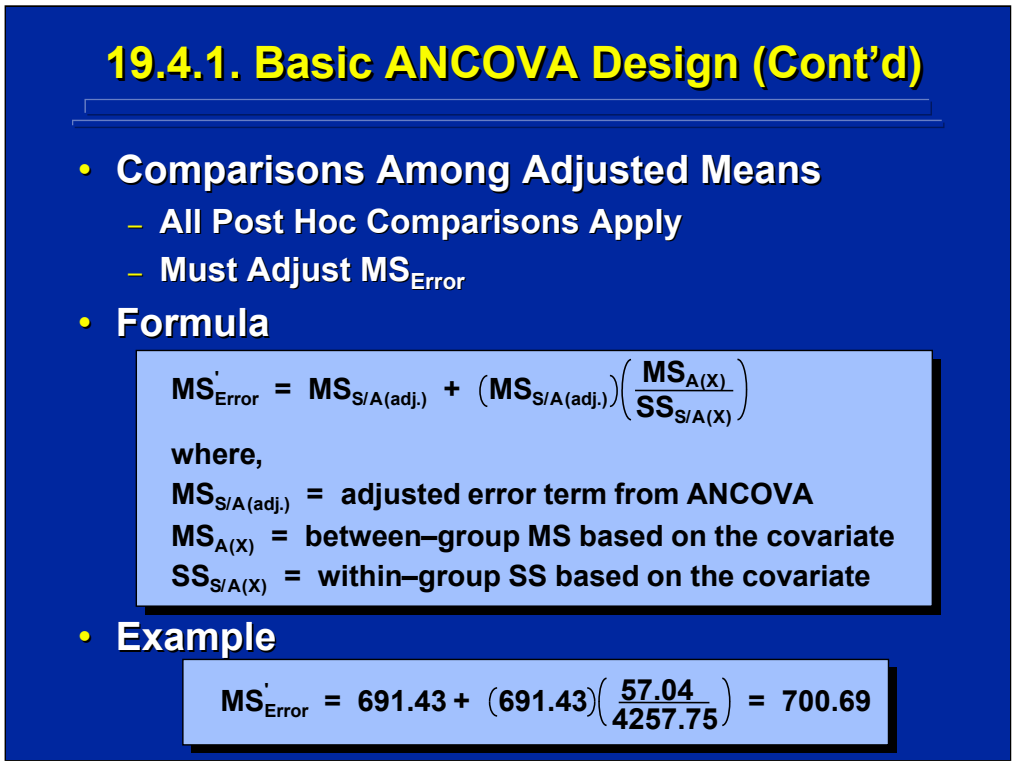
$$\bar{Y}'_{A_1} = 277.50 - 1.61(189.50 - 188.42) = 275.76$$

$$\bar{Y}'_{A_2} = 298.75 - 1.61(185.38 - 188.42) = 303.64$$

$$\bar{Y}'_{A_3} = 273.88 - 1.61(190.38 - 188.42) = 270.72$$

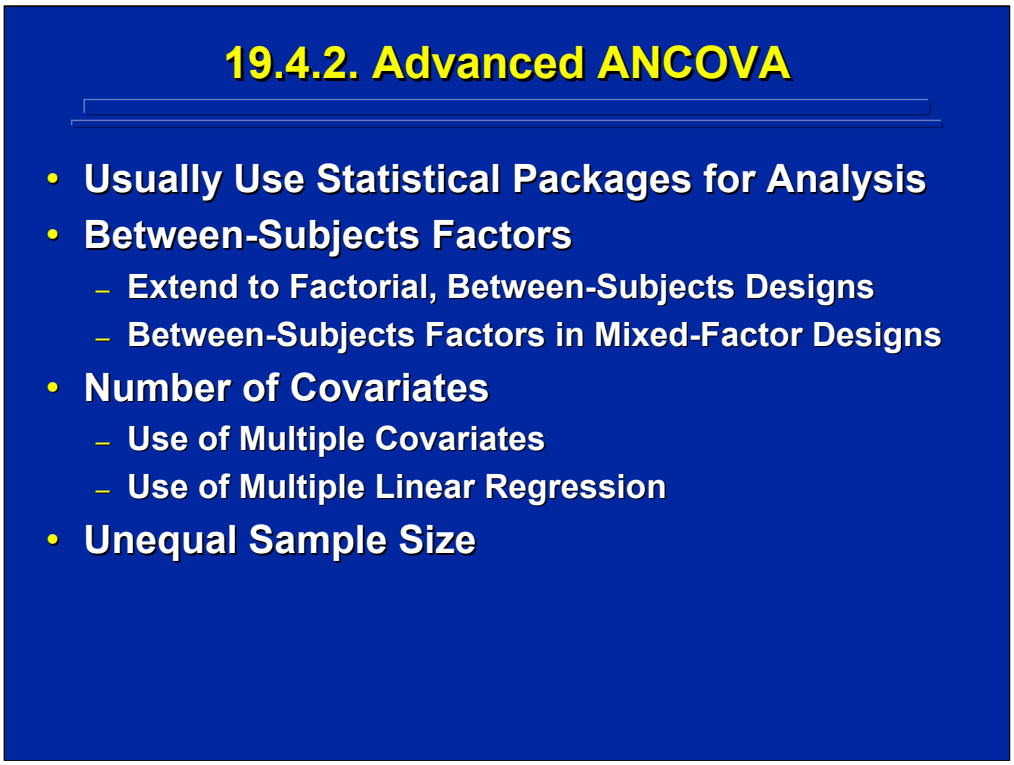
(Click in this red rectangle to see SAS calculations for this example.)

The top portion of this slide shows the formula for adjusting treatment means based on the covariate, X. Examples of using this formula for the three adjusted means in the soldier weight training problem are given in the bottom portion of this slide.




(Click in this red rectangle to see SAS calculations for this example.)

All of the post hoc comparisons techniques described in Topic 11 can be used, but the mean square error must be adjusted for the X covariate, too. The formulae for this adjustment and its use with the example problem are shown on this slide.




Statistical analysis packages are usually used to calculate basic ANCOVA just as Slater and Williges (2006) demonstrated the use of SAS on the weight training example problem. Usually only basic ANCOVA with one covariate is used in human factors research, but several between-subjects factors or between-subjects factors in mixed-factor designs can be considered as long as all treatment means are adjusted for the X covariate. It is also possible to extend basic ANCOVA to include more than one covariate using multiple linear regression, but adjusting treatment means becomes more complex.

Although most human factors experiments have equal sample size to maintain statistical robustness, the ANCOVA can also be used in experiments using unequal sample sizes, if necessary. But, the computations are more complex and need to be weighted by the various sample sizes of the various treatment conditions.

## 19.4.3. Interpretating ANCOVA

- **Comparison of ANOVA to ANCOVA**

| | | ANOVA | | | ANCOVA | |
|---|---|---|---|---|---|---|
| Source | df | SS | F | df | $SS_{adj.}$ | F |
| A | 2 | 2889.25 | 1.22 | 2 | 4912.03 | 3.55* |
| S/A | 21 | 24962.38 | | 20 | 13828.58 | |
| Total | 23 | 27851.63 | | 22 | 18740.61 | |

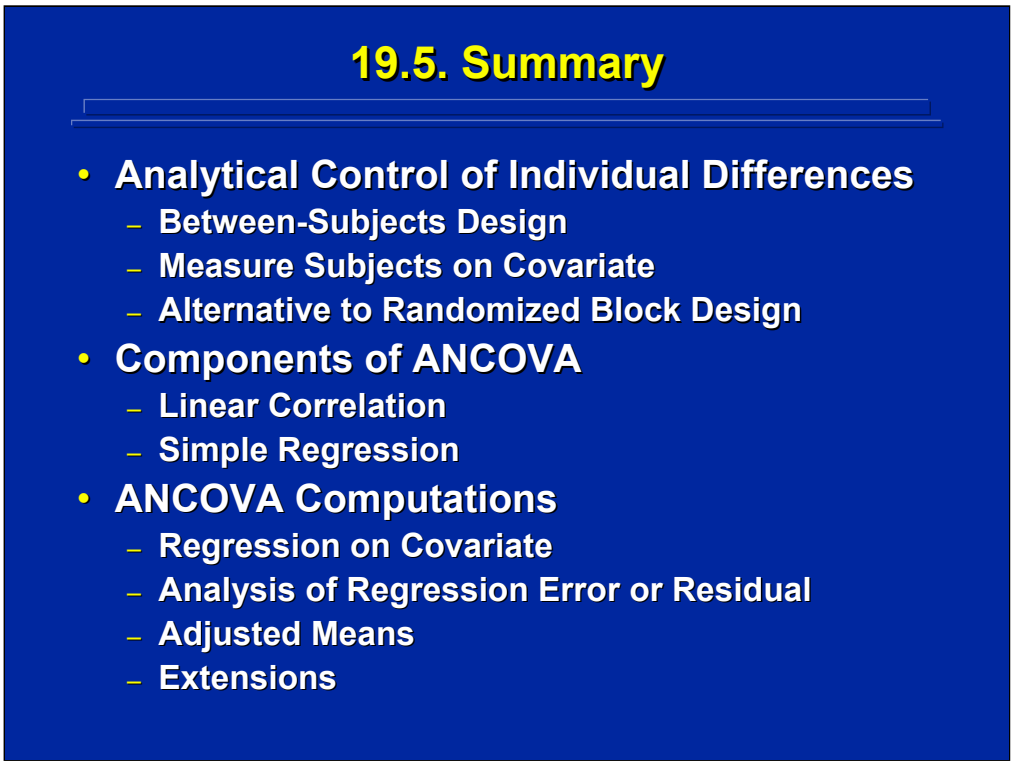
$\bar{Y}_{a_1} = 277.50$ $\quad$ $\bar{Y}'_{a_1} = 275.76$

$\bar{Y}_{a_2} = 298.75$ $\quad$ $\bar{Y}'_{a_2} = 303.64$

$\bar{Y}_{a_3} = 273.88$ $\quad$ $\bar{Y}'_{a_3} = 270.72$

- **Covariate MUST Be Considered When Interpreting Treatment Effects**
  - **Significance is Based on Adjusted Treatments Means**

A major restriction of the ANCOVA is that significant effects must be interpreted in terms of the means adjusted for the X covariate. Both the unadjusted means analyzed in the ANOVA and the adjusted means analyzed in the ANCOVA for the example soldier weight training problem are shown on this slide. The treatment means in the ANCOVA provide a larger spread of means when they are adjusted for the covariate as shown on this slide. Consequently, the significant differences among treatment means in the example problem are significant only when “they are adjusted for with different weights of the soldiers”.




ANCOVA provides an analytical procedure to control for individual differences among subjects in between-subjects designs. Subjects are measured on a covariate that is used to adjust treatment means rather than equally spread the levels of the covariate across treatment levels as done in a randomized block design.

To be successful in making the tests of significance more sensitive, the covariate chosen must be significantly correlated with the dependent variable resulting in a significant simple regression effect (1 df) that is removed from the analysis. Knowledge of linear correlation and simple linear regression concepts reviewed in this topic are prerequisites to ANCOVA. Subsequently, the ANCOVA uses regression error to adjust the treatment means and their error term. Significant results in ANCOVA must be interpreted in term of the means adjusted for the covariate rather than the unadjusted means in ANOVA. Although most applications of ANCOVA involve simple regression, these procedures can be extended to multiple covariates using multiple linear regression and can be extended to adjusting multiple between-subjects factors using equal or unequal sample sizes.

## 19.6. Supplemental Readings

| REFERENCE | SECTION |
|---|---|
| Draper and Smith (1981) | Chapter 1 |
| Hays (1994) | Chapters 14, 15, 17 |
| Hicks & Turner (1999) | Chapter 16 |
| Keppel & Wickens (2004) | Chapter 15 |
| Mason, Gunst, & Hess (2003) | Chapters 14, 15, 16 |
| Maxwell & Dulaney (2000) | Chapter 9 |
| Montgomery (2005) | Chapter 15 |
| Myers (1979) | Chapter 16 |
| Myers (1990) | Chapter 2 |
| Winer, Brown, & Michels (1991) | Chapter 10 |

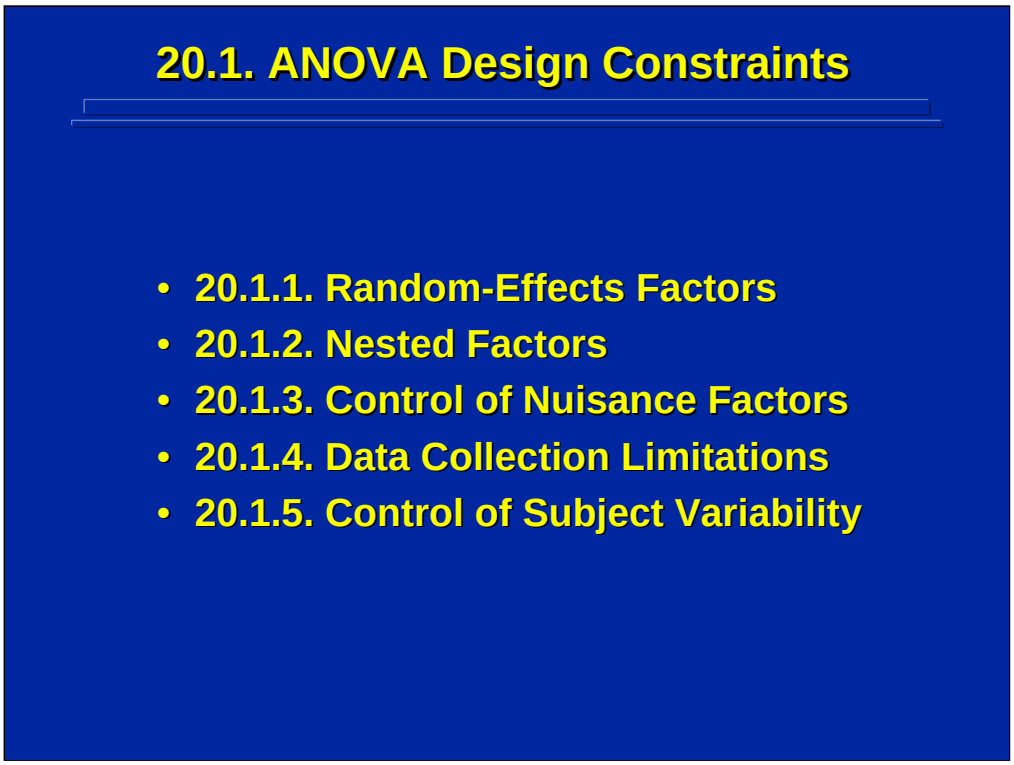
All of these texts provide a discussion of ANCOVA. In addition, Hays (1994) provides a discussion of various correlation techniques and both Myers (1979) and Myers (1990) provide computational details on testing the goodness of fit in simple linear regression.

# Topic 20. Summary of Advanced ANOVA

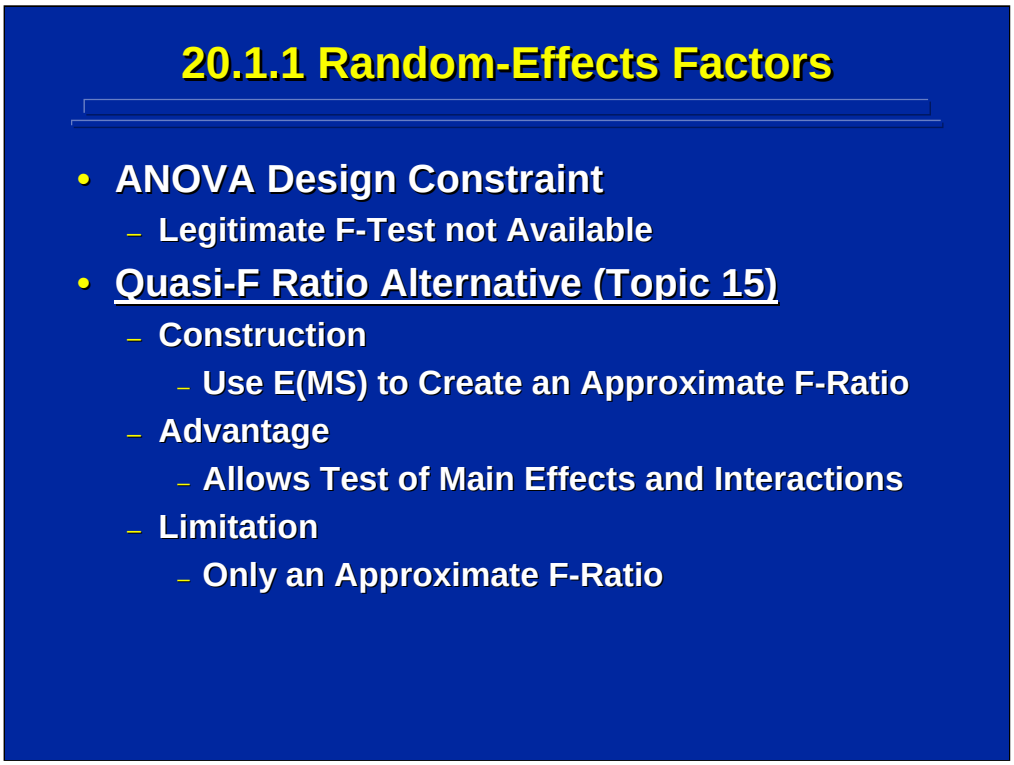
20.1. ANOVA Design Constraints
- 20.1.1. Random-Effects Factors
- 20.1.2. Nested Factors
- 20.1.3. Control of Nuisance Factors
- 20.1.4. Data Collection Limitations
- 20.1.5. Control of Subject Variability

20.2. Advanced ANOVA Design Process
20.3. ANOVA of Regression Analysis
20.4. Summary
20.5. Supplemental Readings

To summarize the advanced ANOVA techniques described in Section 4, composite of design constraints and a process for addressing them is presented in this topic. The use of ANOVA in regression analysis that was introduced in Section 4 will be expanded in Section 5. This topic ends with a summary and a complete listing of supplemental readings covered in Section 4.

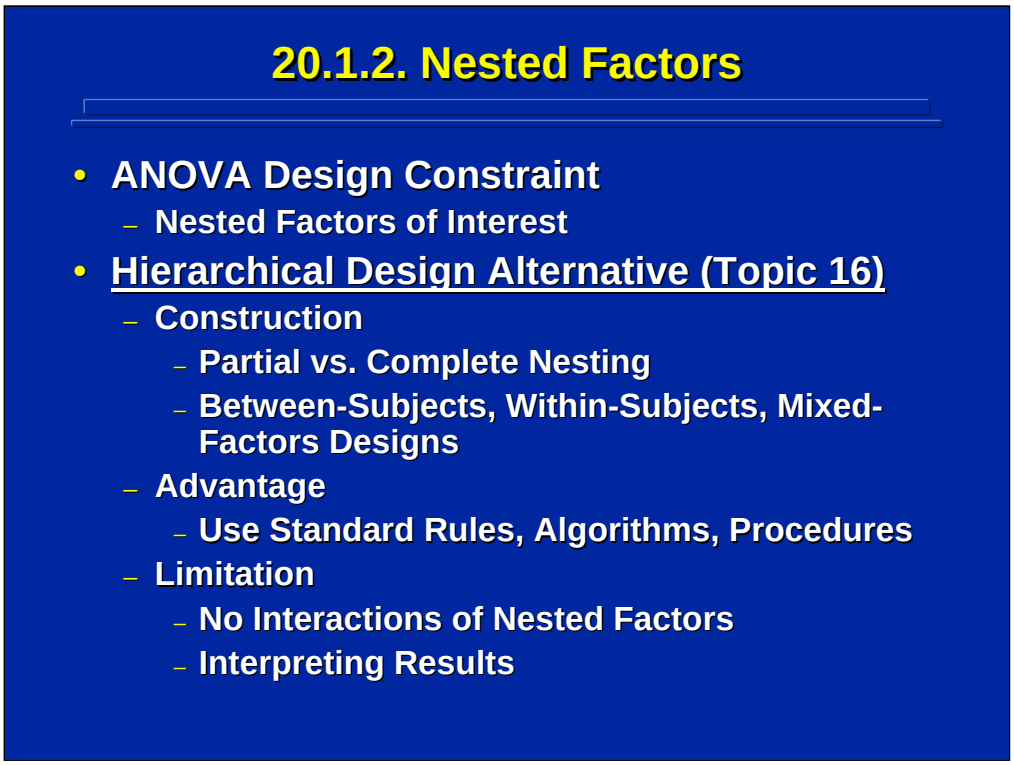



Advanced ANOVA designs primarily deal with experimental design constraints that require extensions of basic ANOVA design procedures. Five major experimental design constraints that often occur in the human factors and ergonomics research are covered in Section 4. These five constraints listed on this slide are discussed separately in this subsection.




When some of the factors of interest to the experimenter exist only as random-effects variables, legitimate F-ratios may not exist to test these effects. Quasi-F ratios described in Topic 15 can be used if this constraint occurs.

Quasi-F ratios are based on E(MS) and required various MS quantities to be added and subtracted together. Since the resulting ratio only approximates an F-ratio, the tabled value of F is adjusted by the Satterwaithe correction for df. The quasi-F allows the experimenter to test various main effects and interactions of random effects factors if approximate F-ratios are acceptable to the experimenter.




At times, some or all the factors of interest in human factors research are nested rather than crossed. In this situation hierarchical ANOVA designs rather than basic, completely crossed factorial designs are needed as described in Topic 16.

Depending on the nesting relationships among factors, either partial or complete hierarchical designs need to be constructed. Using the standard rules, algorithms, and procedures of basic ANOVA, hierarchical designs can be constructed as between-subjects, within-subjects, or mixed-factors designs depending on the assignment of subjects to treatment conditions. Since nested factors cannot interact, certain interactions do not exist in hierarchical designs. In addition, the interpretation of main effects of nested factors becomes problematic.

## 20.1.3. Control of Nuisance Factors

- ANOVA Design Constraint
  - Nuisance Variables (e.g., Multiple Data Collection Sessions, Multiple Experimenters) Confounding
- Blocking Design Alternative (Topic 17)
  - Construction
    - Modular Representation
    - Simple vs. Complex Blocking
    - Defining Relationship
  - Advantage
    - Limits Confounding Effect of Nuisance Variable
  - Limitation
    - Equal Levels of Each Factor

In complex ANOVA designs requiring data collection across sessions or using more that one experimenter, the effects of multiple sessions and experimenters become confounded with the factors of interest to the experiment. The confounding of these so-called nuisance variables can be controlled through the use of blocking designs described in Topic 17.

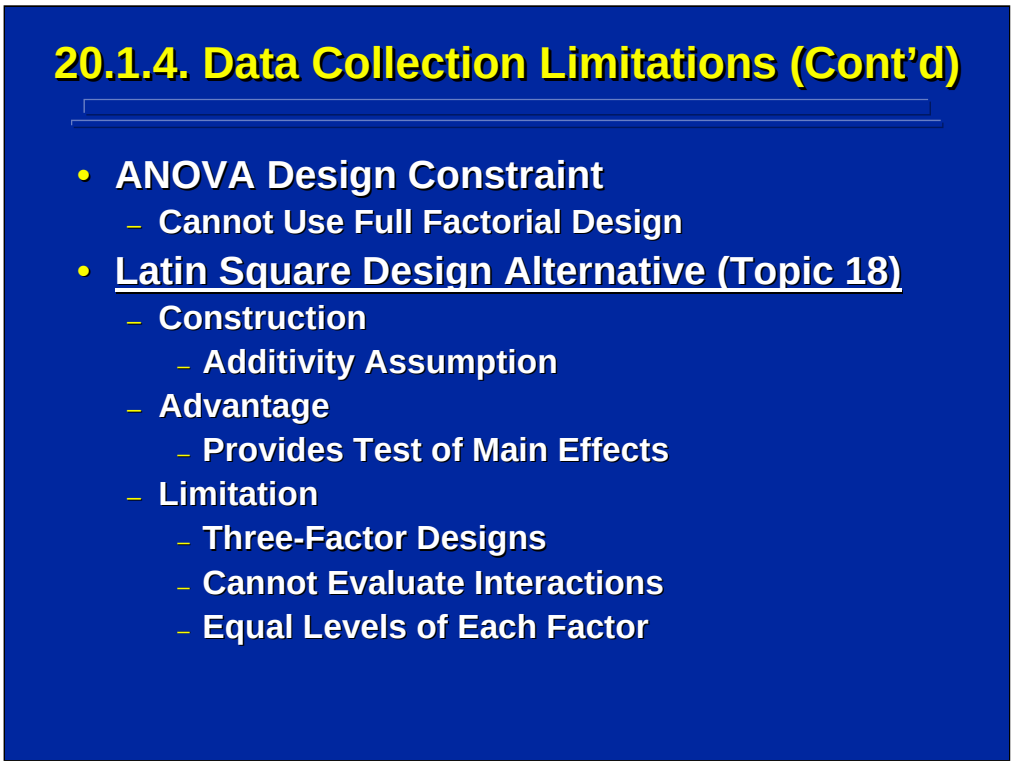
A defining relationship described in modular notation can be used to specify the confounding of effects of interest with a nuisance variable in simple blocking. Multiple defining relationships and their generalized interactions are confounded in complex blocking. By using blocking procedures, the experimenter can keep main effects and two-way interactions unconfounded with the nuisance variable. These blocking procedures, however, require that each factor in the blocked design have the same number of levels.

## 20.1.4. Data Collection Limitations

- ANOVA Design Constraint
  - Cannot Use Full Factorial Design
- $2^{k-p}$ Fractional-Factorial Design Alternative (Topic 18)
  - Construction
    - Modular Representation
    - Identity Relationship
    - Alias Structure
  - Advantage
    - Design Resolution
  - Limitation
    - Equal Levels of Each Factor

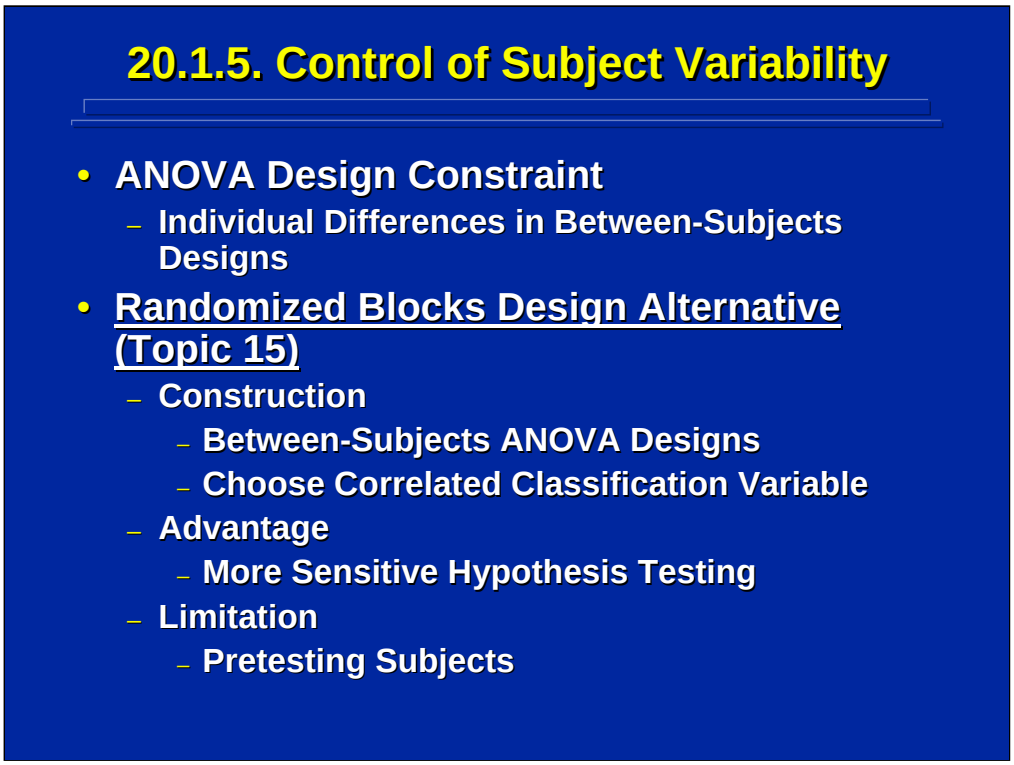
Time and money constraints may not allow collection of all the necessary data in a full factorial design that involves a large number of factors. A fractional-factorial design can be considered as an alternative in these situations.

Topic 18 describes $2^{k-p}$ fractional replicates as a useful fractional-factorial design for human factors research. Modular representation is used to construct the fractional-factorial designs such that the identity relationship defines the effect or effects that cannot be evaluated and the alias structure lists the confounded effects in the $2^{k-p}$ fractional replicate. The experimenter can keep all main effects and two-way interactions unconfounded in the $2^k$ factorial design if the experimenter chooses a Resolution V fractional replicate. Fractional replicates require an equal number of levels of each factor. Besides $2^k$ designs, fractional factorials can also be constructed for $3^k$ and $5^k$ factorial designs using these procedures if the experimenter can accept partial effect confounding.



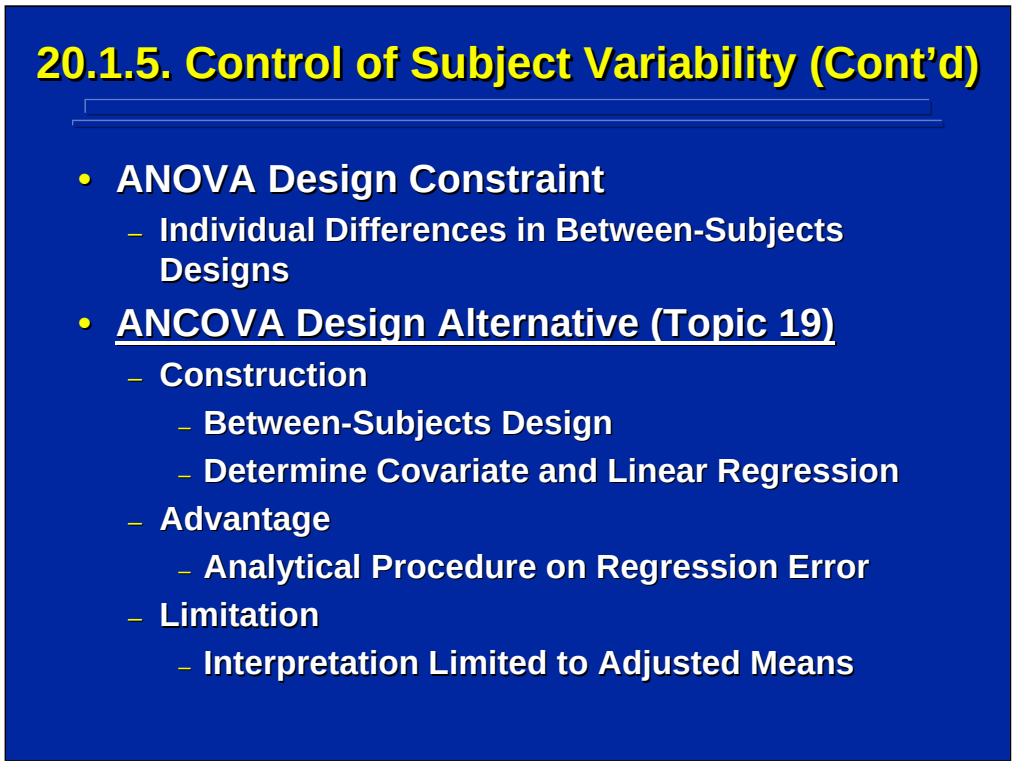

When a fractional-factorial design with three factors is required, a Latin square design is useful to evaluate just the main effects of the three factors. This special case of a fractional-factorial design is described in Topic 18.

All Latin square designs consist of three factors each with the same number of levels of each factor. The data matrix appears as a square matrix with rows and columns defined by the factor levels of two factors, and each level of the third factor appears once within the rows and columns of the data matrix. The experimenter assumes additivity of the three factors, meaning no interactions exist in order to use the residual as the error term in conducting F-tests. Only main effects of the three factors can be evaluated in Latin square designs.




The difference between subjects is one the main source of variability in human factors research. This variability can make between-subjects designs less sensitive to testing effects of interest. Topic 15 describes a method for removing between subjects variability through a randomized blocks design.

Subject classification variables such as gender and experience that have known correlations with the dependent variable are usually chosen as the blocking variable in these designs. An equal number of subjects at each level of the classification variable are randomly assigned to each treatment condition in the between-subjects design of interest. Subsequently, the effect of the blocking variable is removed from the error term in the ANOVA to result in more sensitive F-tests on the factors of interest. These designs can be quite useful in removing sources of between subjects variability in between-subjects designs, but a randomized blocks design requires additional effort in pretesting subjects on the classification variable.




An alternative to the randomized block design is to remove individual differences in between-subjects designs analytically through ANCOVA. This analytical procedure is described in Topic 20.

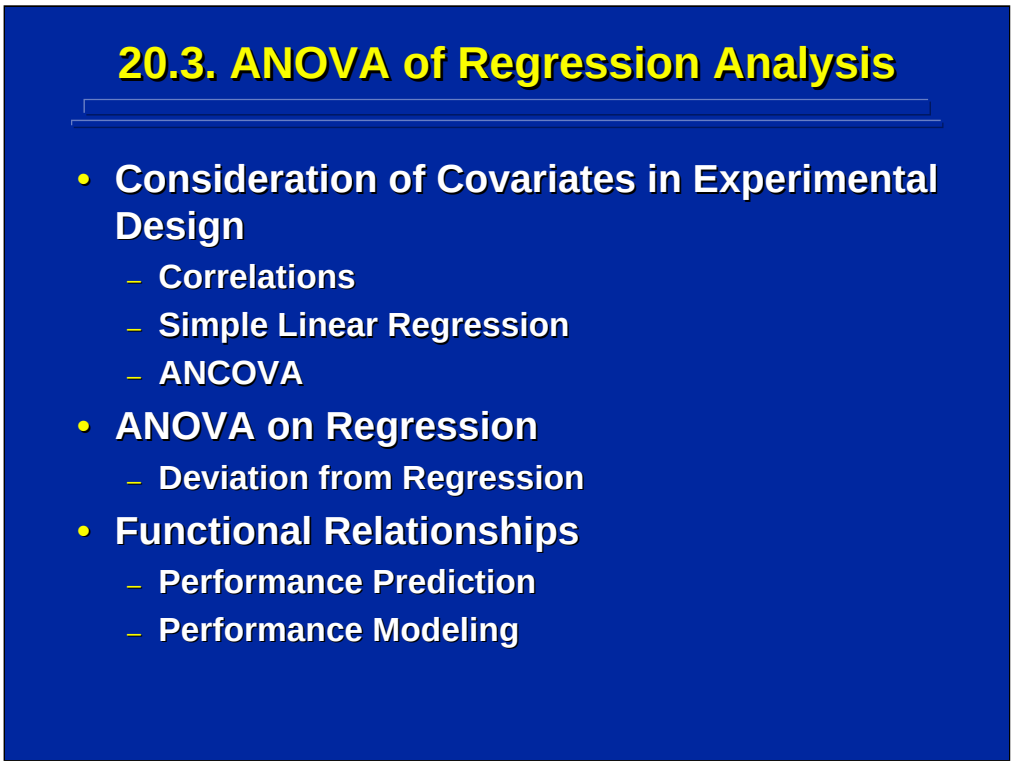
The subject classification variable that covaries with the dependent variable in the experiment is evaluated using simple linear regression. The residual or regression error is then used in the ANCOVA to evaluate the effects of interest that are adjusted for the effect of the covariate. This analytical procedure is straightforward and provides a useful technique for removing individual difference effects from the experiment as long as the experimenter is willing to interpret the results in terms of means adjusted for the covariate.

## 20.2. Advanced ANOVA Design Process

- Step 1. Consider Real-World Constraints
  - List Possible Constraints
- Step 2. Consider ANOVA Design Alternatives
  - List Candidate Design Alternatives
  - Select Candidate Alternatives
- Step 3. Trade-off ANOVA Design Alternatives
  - Consider Advantages of Possible Design Alternatives
  - Consider Limitations of Possible Design Alternatives
- Step 4. Choose Appropriate ANOVA Design
  - Implement Advanced ANOVA Design Procedure

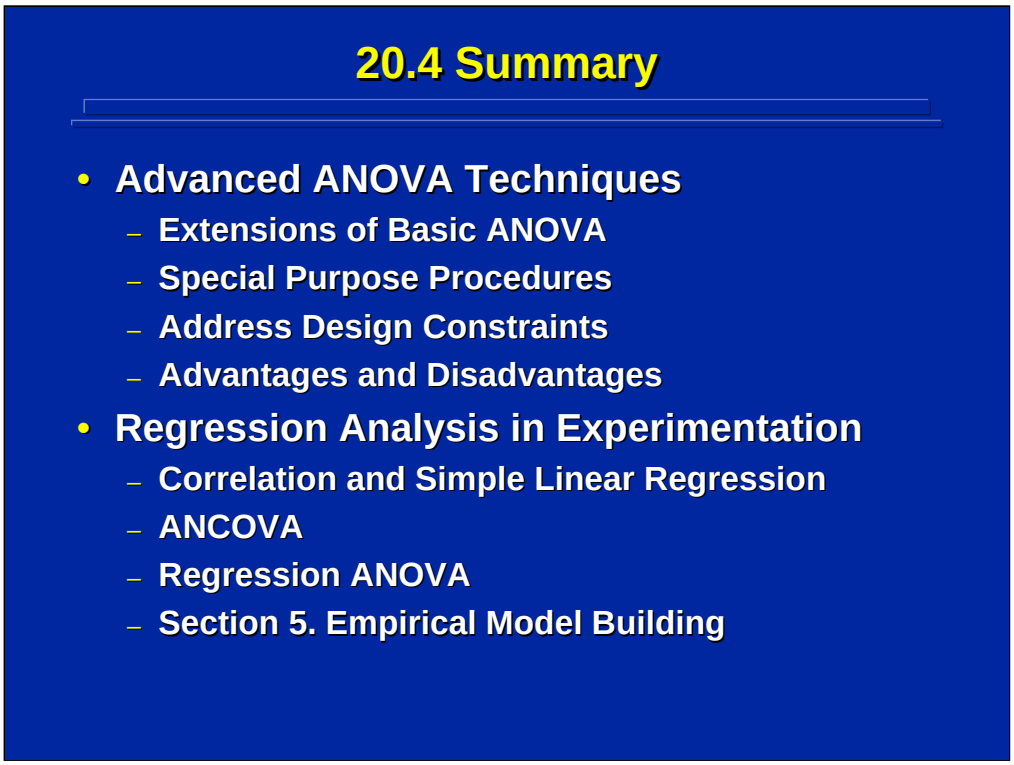
This slide summarizes the four step process that can be used to choose the appropriate advanced ANOVA procedure described in Section 4. In Step 1, the real-world constraints of the experiment must be listed. Once these constraints are known, candidate advanced ANOVA techniques are considered as noted in Step 2.

Viable alternatives resulting from Step 2 are evaluated in Step 3. For example, the use of randomized block designs and ANCOVA can be considered as a means of minimizing the effect of individual differences in between-subjects designs. After considering the various trade-offs in Step 3, the appropriate advanced ANOVA procedure is selected and implemented in Step 4.



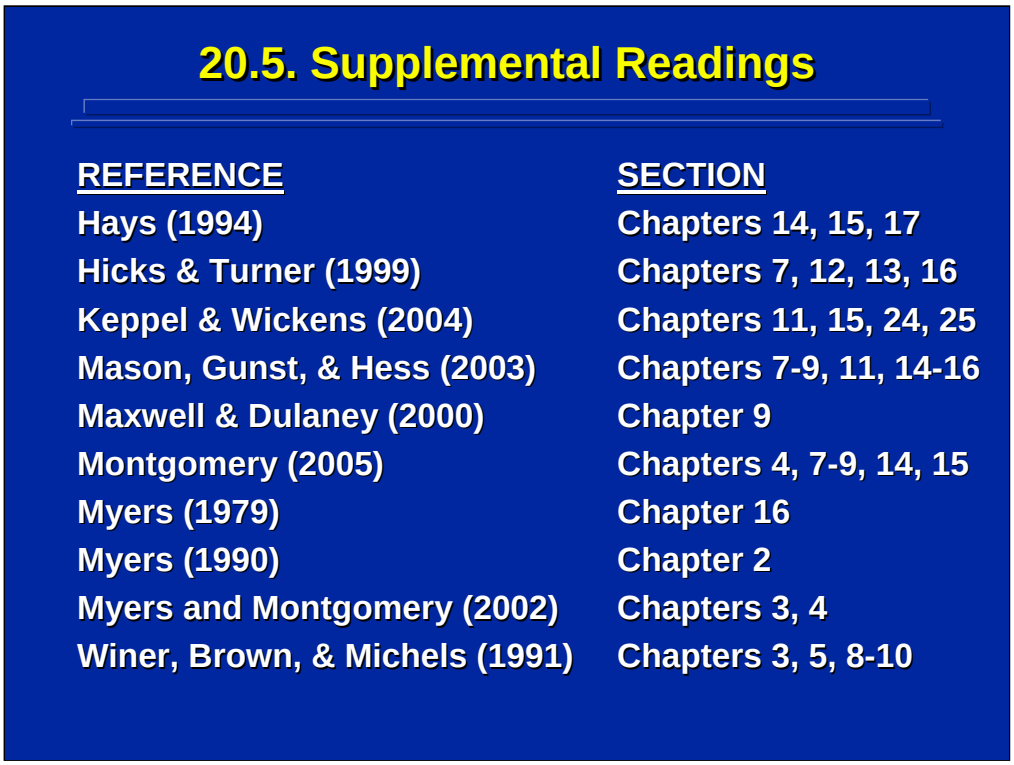

Topic 19 introduced the concept of considering covariates in experimental design. These covariates are correlated with the dependent variable of the experiment. Concepts of correlation and simple regression form the basis of ANCOVA procedures described in this topic.

More importantly, information covered in Topic 19 demonstrated that the regression analysis can be analyzed by conducting an ANOVA on a general linear model. This ANOVA evaluates the significance of the regression model based on a least squares criterion by using the deviation from regression as the error. Subsequently, partial F-tests can be conducted on the parameters of the linear model. Simple regression is an empirical model that predicts performance as a function of one predictor.

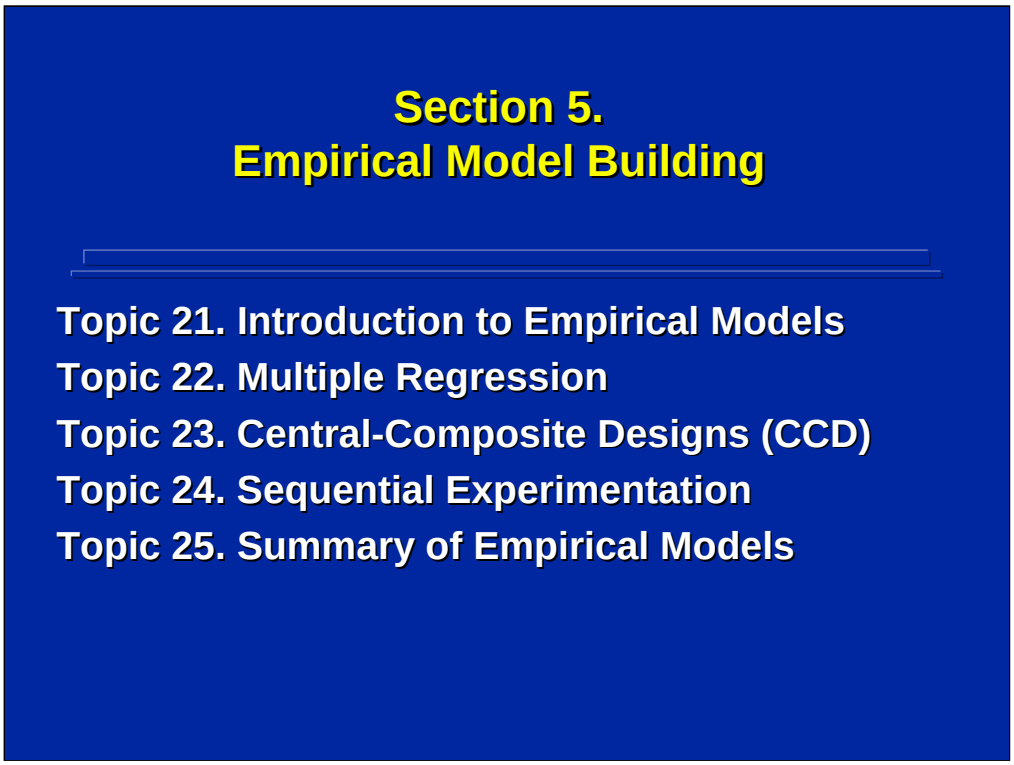



By way of summary, this section on advanced experimental designs is simply an extension of basic ANOVA that addresses various real-world experimental design constraints. The advantages and disadvantages of several special purpose procedures are discussed in this section to aid the experimenter in choosing the appropriate design alternative.

Simple linear regression was introduced as a technique used in ANCOVA to remove the effects of individual difference that are correlated with the dependent variable in the experiment. An ANOVA can be conducted using the linear regression model. In Section 5, the use of regression analysis in experimentation is extended to experimental design techniques that are useful in building empirical models that predict performance as a function of several independent variables.




This slide provides a summary of supplemental reading chapters on all the topics presented in Section 4. Advanced ANOVA. The Hays (1994), Myers (1979), and Myers (1990) chapters primarily provide a basic review of correlation and simple regression. Montgomery (2005) is a non-behavioral science experimental design textbook, and the chapters in the remaining texts deal primarily with behavioral science research applications using various advanced ANOVA experimental designs.




Section 5 is the last section of the human factors reference material and incorporates material presented in the other sections. The emphasis in this section is on the use of experimental design and analysis procedures to build second-order, empirical models that predict human performance in complex systems applications. This section covers the following topics:

Topic 21 – an introduction to empirical models;

Topic 22 – multiple linear and polynomial regression;

Topic 23 – second-order empirical model building using central-composite designs;

Topic 24 – response surface exploration and sequential experimentation; and

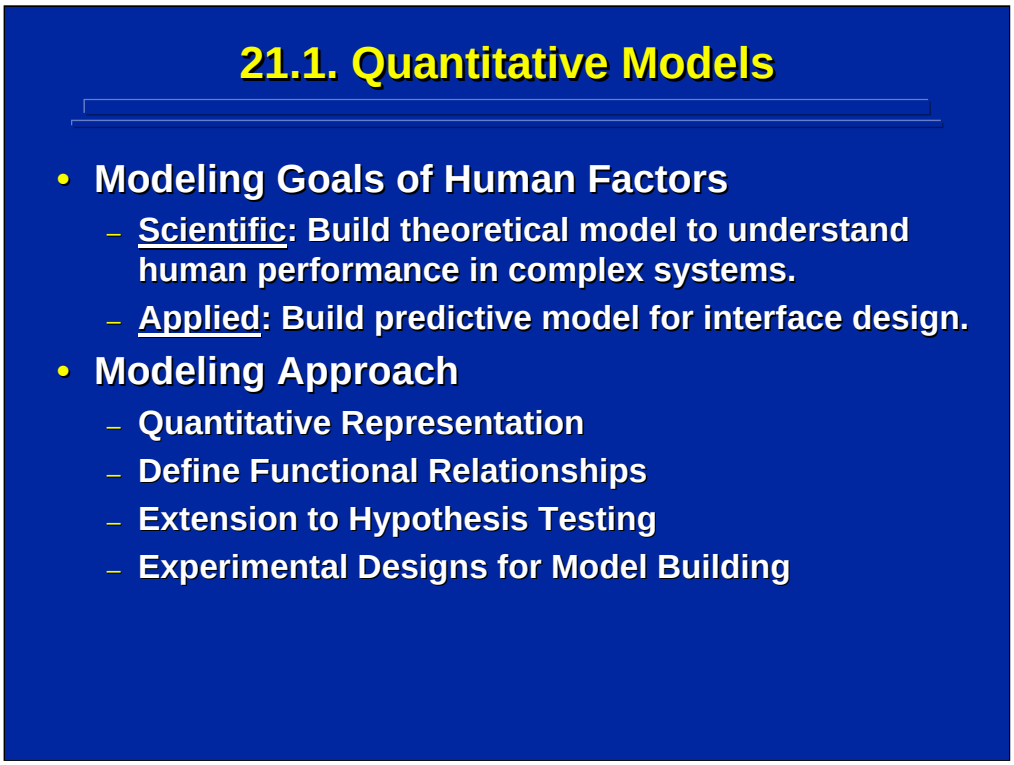
Topic 25 – a summary of empirical models and overall conclusions.

# Topic 21. Introduction to Empirical Models

21.1. Quantitative Models
- 21.1.1. Mechanistic Models
- 21.1.2. Empirical Models

21.2. Model Building Experiments
21.3. Models in Human Factors
21.4. Summary
21.5. Supplemental Readings

This introduction provides an overview of quantitative models with an emphasis on empirical model building developed through efficient experimental design. Empirical models are descriptive models of human behavior based on results obtained through one or more controlled experiments that can be used to predict human performance in complex systems. These empirical models can assist the human factors specialist in conducting design tradeoffs of critical interface parameters. A summary of the topics covered in Section 5 that support empirical model building is provided along with supplemental readings on the general topic of quantitative models.




Quantitative modeling in human factors research can have scientific and applied goals. When the goal is primarily scientific, the human factors specialist is interested in building a theoretical model to aid in understanding human performance in complex systems. When the goal is applied, the human factors specialist is interested in building a model that predicts actual performance in a specific interface context. This section describes a statistical approach that results in empirical models.

Both goals of modeling can be represented quantitatively in terms of a prediction equation of human performance as a function of the weighted influence of critical independent variables. The resulting functional relationship is an extension of hypothesis testing which tests only the statistical significance of independent variables in an experiment. Experimental designs can be used as an efficient way of collecting the necessary and sufficient data to build quantitative models of human performance.

## 21.1. Quantitative Models (Cont'd)

- **General Form**

$$\eta = f(\beta, X)$$

where,

$\eta$ = Expected, Predicted Outcome i.e., E(Y)

$\beta$ = Defining Parameters of the Situation

X = Independent Variables Affecting Outcomes

- **Human Factors Applications**

where,

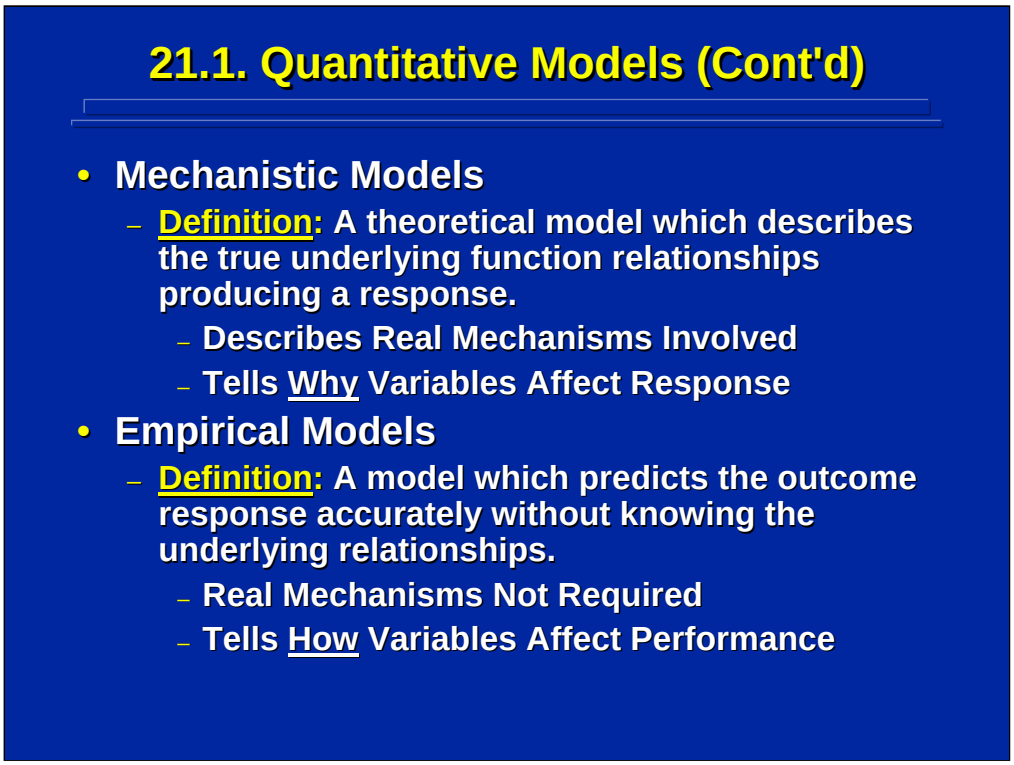
$\eta$ = Operator's Expected, Predicted Performance

$\beta$ = Task Parameters

X = Independent Variables Including Human, Machine, and Environment

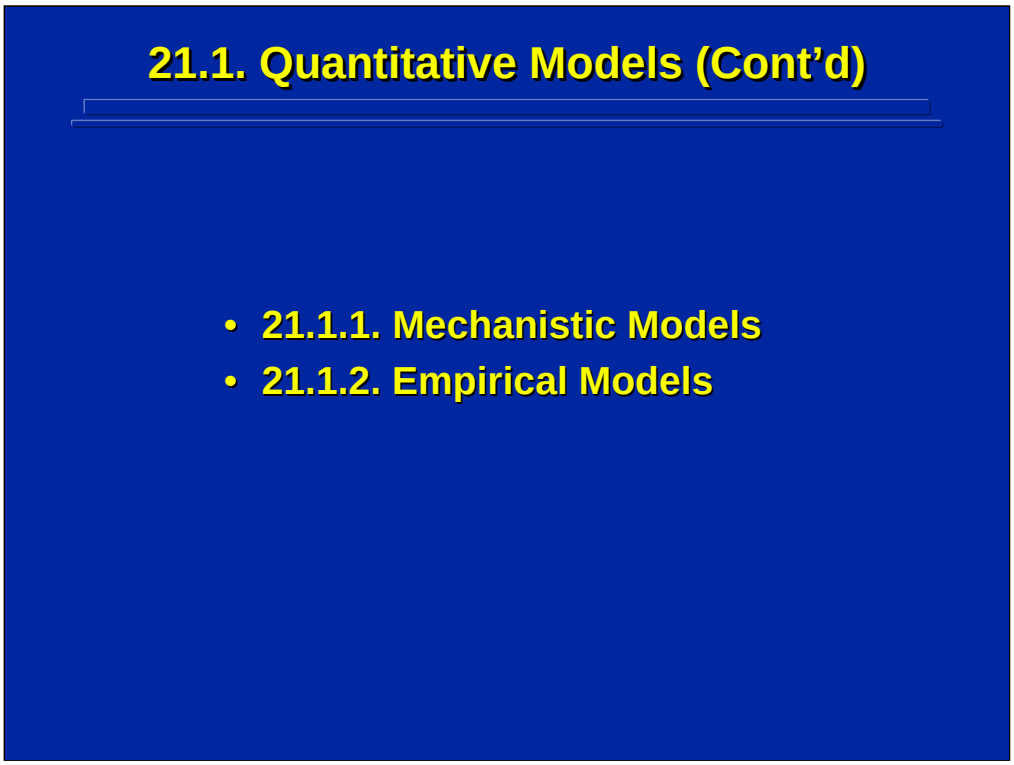
The general form of any quantitative model of the expected value of an outcome, Y, as a function of independent variables, Xs, is shown on this slide. In general, the predicted outcome, $\eta$, is a function of X weighted by specific defining variables of the situation, $\beta$.

In human factors research, $\eta$ is the dependent variable or the predicted performance. The values of $\beta$ are the parameters in the functional relationship that defines a particular task situation. These parameters, in turn, weight the various independent variables, Xs, that are aspects of the human, machine, and environment interface.

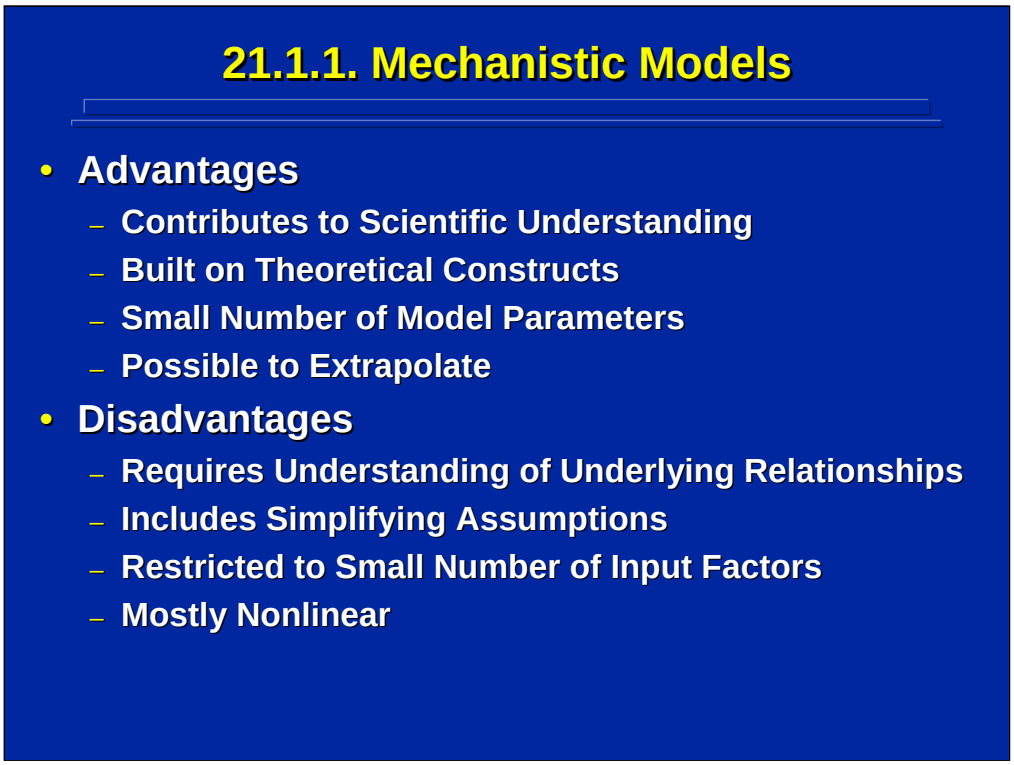



The independent variables in quantitative models can be described either in terms of a mechanistic model or an empirical model. The mechanistic model is the theoretical model which describes the true underlying relationships producing a response. The goal of this model is primarily to advance scientific understanding of why variables in the model affect performance based on the effect of underlying mechanisms such as the laws of physics.

Empirical models, on the other hand, predict the outcome response as a function of situational variables without knowing the true underlying relationships or mechanisms. This type of quantitative model predicts how much each variable affects performance, not why each affects performance. Often empirical models are used as a starting point to develop theoretical or mechanistic models.

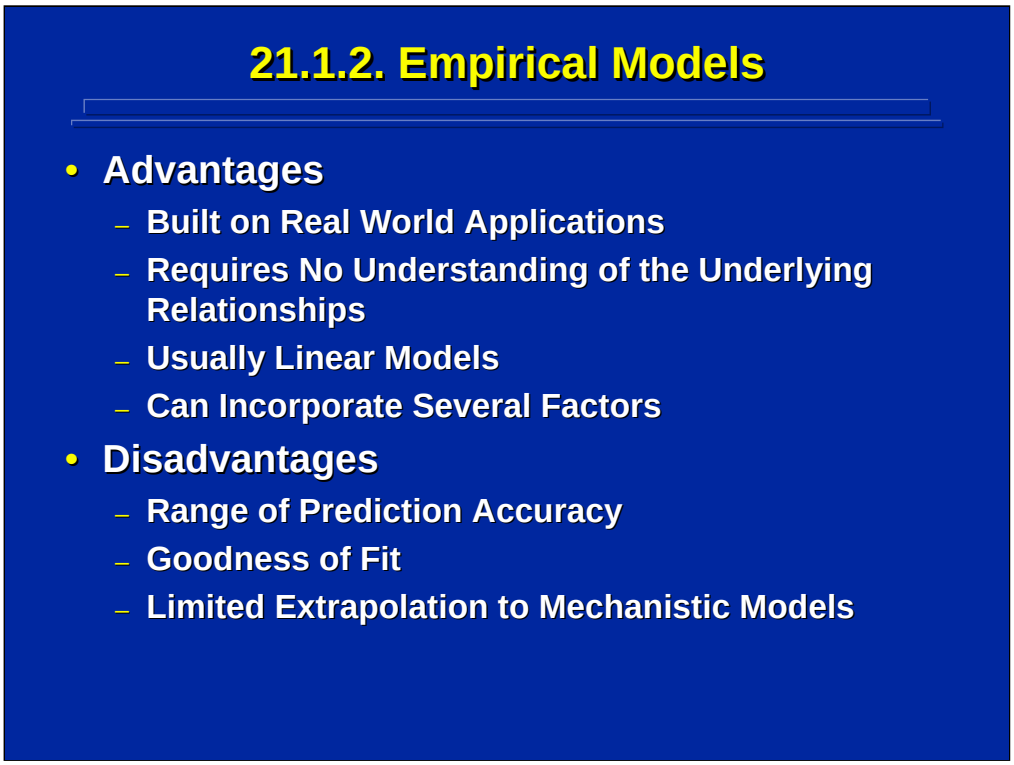



This subsection provides some elaboration on the distinction of mechanistic and empirical models following the distinctions made by Box and Draper (1987) and Box, Hunter, and Hunter (2005). The emphasis of this reference material is on the development of empirical models that can be used in human factors and ergonomics interface design.




The advantage of the mechanistic model is that it contributes to scientific understanding and is based on established scientific constructs. Box and Draper (1987) note that mechanistic models can be extrapolated across the range of input variables. They are also parsimonious since they include only a few parameters.

The major disadvantages of theoretical models are that the researcher must have a good understanding of the underlying relationships in order to build them, and mechanistic models cannot handle complex relationships economically without simplifying assumptions. Therefore, applications may be limited to a specific, small number of input factors. In addition, mechanistic models are often nonlinear requiring more complex mathematical treatment.




Empirical models are built on data drawn from real-world applications and do not require a detailed understanding of the underlying relationships. Most empirical models are linear and based on least squares regression procedures. These procedures can handle complex relationships involving many factors in an economical way and are quite useful in human factors applications.

Prediction accuracy is limited to the range of the factor levels observed in generating the empirical model. So, careful attention must be given to sampling the appropriate range of interest and not using the empirical model for prediction beyond those ranges. Goodness of fit of the empirical model must be adjusted for the number of factors included to avoid inflated model validation. Currently, extrapolation procedures to evolve empirical models into mechanistic models are limited. Empirical models can aid in developing an understanding of underlying relationships and possibly lead to mechanistic model development. Box, Hunter, and Hunter (2005, pp. 518-526) provide an example of this type of extrapolation in the chemical sciences.

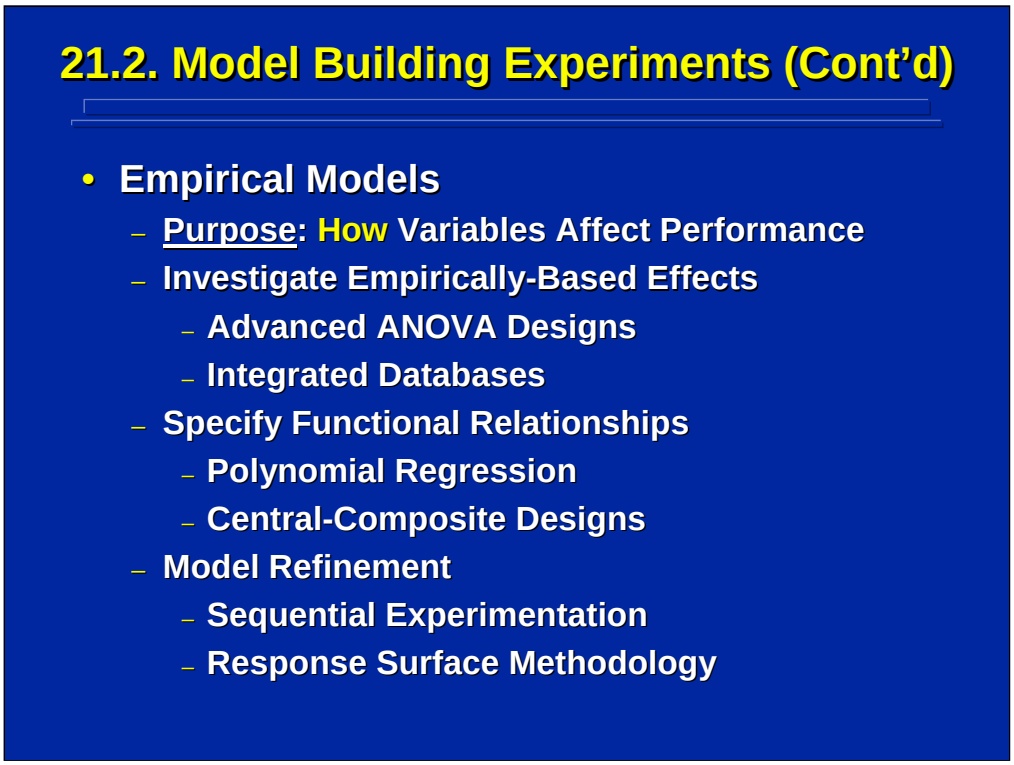
## 21.2. Model Building Experiments

- Mechanistic Models
  - Purpose: Why Variables Affect Performance
  - Model Development
    - Parameter Estimation
    - Model Description
  - Model Testing
- Screening Variables
  - Purpose: Which Variables to Investigate
  - Reduce List of Potential Factors to Most Influential Factors
  - Economical Experimental Designs
  - Main Effects and Two-Way Interactions

The primary goal of Section 5 is to demonstrate the use of experimental designs to develop quantitative models through experimentation. In a more general sense, Box and Draper (2005, pp. 10-14) describe how experimentation can be used to evaluate mechanistic models, screen variables, and build empirical models.

Experiments can be used to facilitate parameter estimation and alternative forms for describing mechanistic models which describe why variables affect performance. Most often experiments are used to test the limits of mechanistic models and improve them.

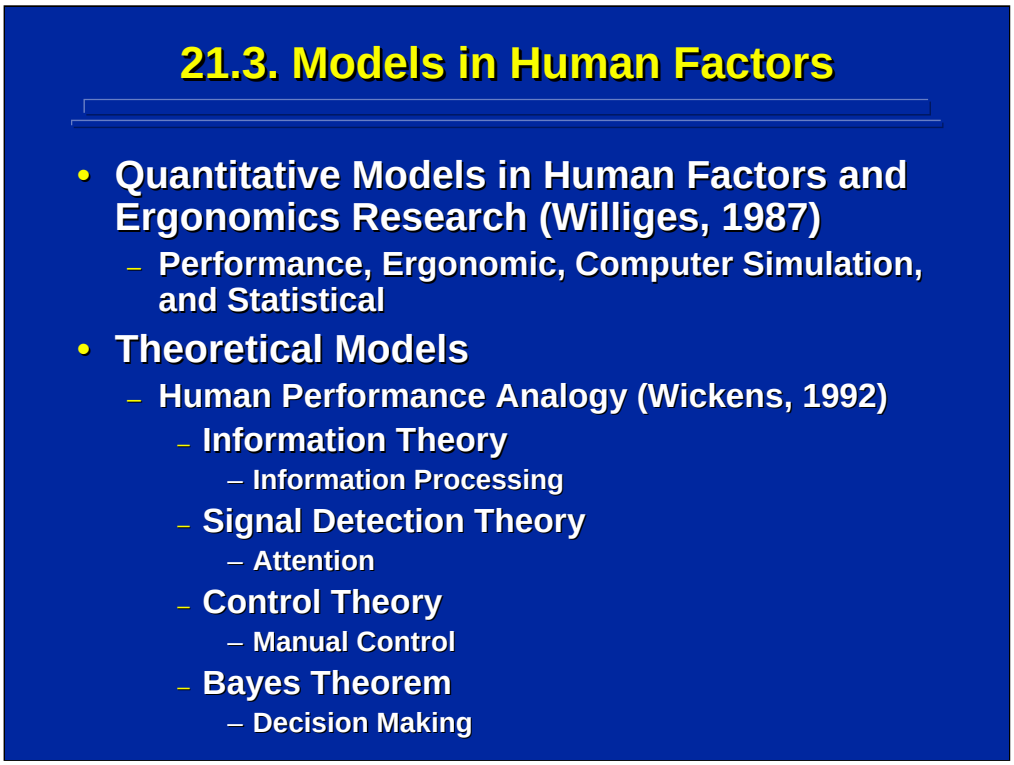
Screening experiments are used to determine which variables need to be included in models. This is a necessary first step in building empirical models because the initial set of potential variables can be quite large and must be narrowed to a reduced set of the most influential variables. Economical data collection experimental design such as single observation factorial designs and fractional replicates can be used for this purpose. In human factors research, the focus of these experiments is to screen variables in terms of main effects and two-way interactions. Higher-order effects are usually of minor interest.




Section 5 focuses on experimental designs that are useful in developing and using empirical models in human factors research. These empirical models specify functional relationships of how variables affect performance. Advanced ANOVA techniques are used to determine the database of empirical effects to model. Experiments are used to screen variables, and sequential experiments are used to build integrated databases.

Polynomial regression is used as a convenient form of empirical models to specify the functional relationship of how several factors affect performance. Efficient, second-order experimental designs such as central-composite designs are used to collect the data for inclusion in polynomial regression prediction equations. The resulting empirical models can be evaluated through ANOVA procedures.

Sequential experimentation is used to refine empirical models by conducting an integrated set of small experiments that can be combined into an integrated database. These procedures are drawn from response surface methodology techniques which allow description and exploration of performance effects specified by empirical models.




Williges (1987) describes quantitative models in human factors and ergonomics research used to predict user performance in human-computer interface design. These modeling efforts use a variety of approaches including observed performance to provide procedural representations, ergonomics-based data to provided anthropometric and biomechanical representations, computer simulations to provide task sequence representations, and statistical-based data to predict human performance.

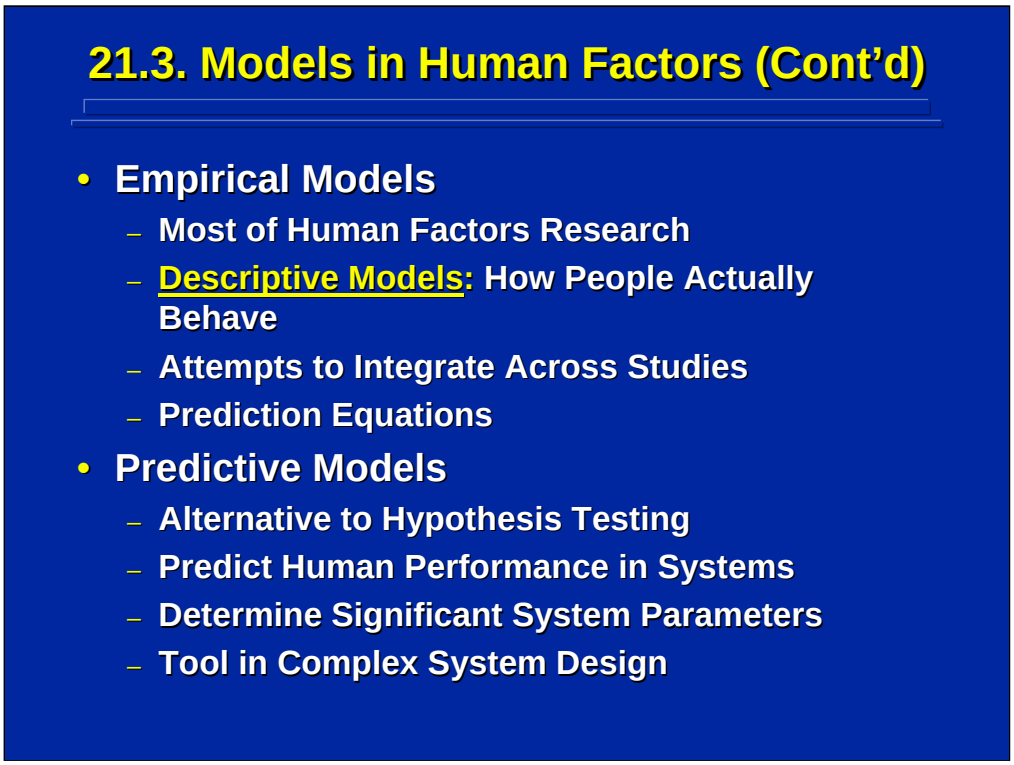
There are some examples of mechanistic models in human factors, but these quantitative models are borrowed from other disciplines. The four examples of theoretical human performance models discussed by Wickens (1992) shown on this slide are borrowed from engineering and probability theory. Information theory has been used to evaluate human information processing; signal detection theory has been used in modeling human attention in vigilance applications; closed-loop control theory has been used in modeling human manual control; and Bayes theorem has been used to model the integration of information in human decision making.

## 21.3. Models in Human Factors (Cont'd)

- Normative Models in Human Factors
  - Normative Models: How People Ought to Behave
  - Major Use of Existing Theoretical Models
  - Usefulness of Normative Models
    - Logical Structuring of Task
    - Suggest Variables and Methods
    - Standards to Evaluate Performance
    - Common Metrics for Human and Machines
    - Revise to Descriptive Models (e.g., Hick-Hyman Law, and Fitt's Law)

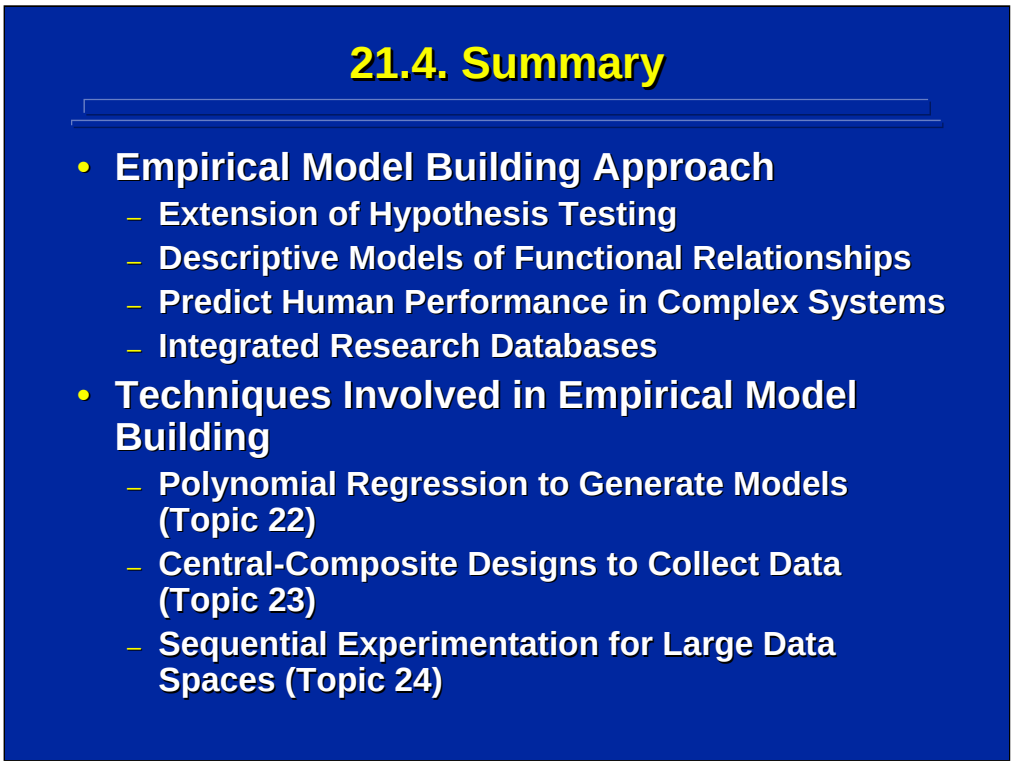
The four theoretical models listed on the previous slide are normative models of human performance, meaning that they specify how people ought to behave. They are not, necessarily, good descriptive models of how people actually behave. Most existing mechanistic models in human factors are normative rather than descriptive models of human performance.

Several important uses of normative models are listed on this slide. The parameters of normative models can be used to structure the interface and provide a list of variables to investigate through experiments as well as methods to investigate them. The normative value predicted by the theoretical model provides a standard of optimal performance for comparison to actual human performance. It may be possible to revise the normative model into a true descriptive model within certain constraints. Hicks-Hyman Law (Wickens, 1992, pp. 317-18, and p.323) and Fitt's Law (Wickens, 1992, pp. 446-449 and pp. 482-483) are two example of descriptive models based on Information Theory.



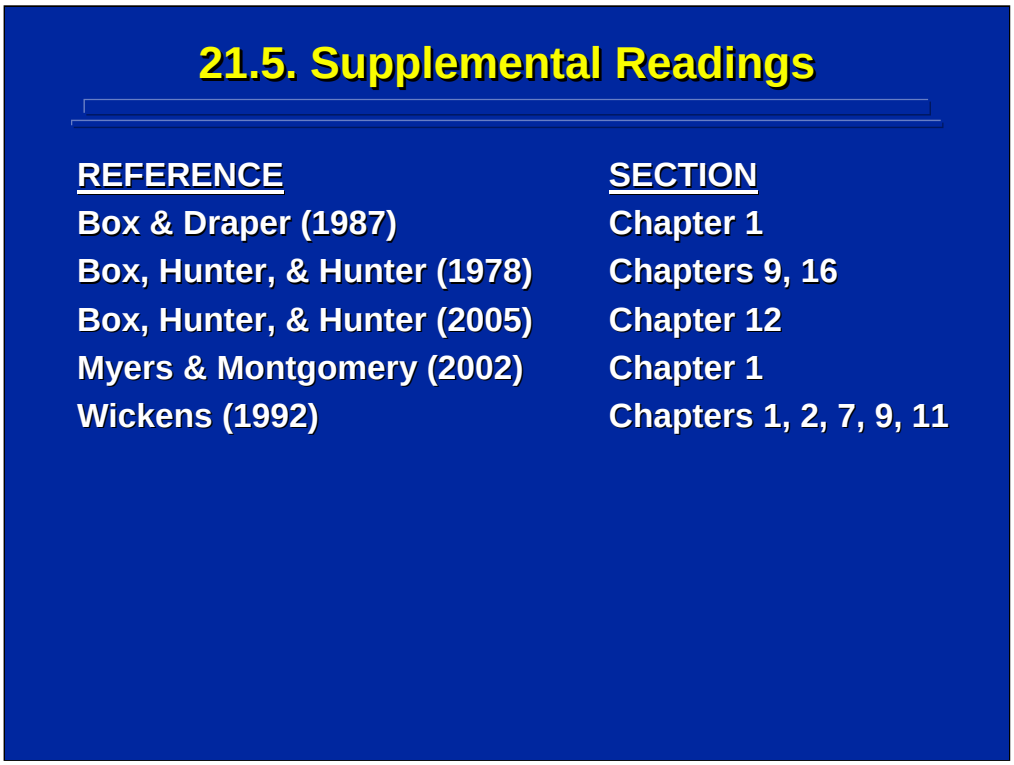

Most quantitative models in human factors are models based on a body of empirical data gathered through a series of experiments that is focused on understanding how people actually behave in complex systems. These empirical models are usually specified in terms of predictions of human performance as a function of task parameters.

Predictive models of actual performance go beyond hypothesis testing of single parameters to evaluate the relative weightings of several parameters in predicting human performance. These relative weightings can be used for interface design tradeoffs and the determination of the most important system parameters to consider in design. So, properly developed and used empirical models can prove to be an important system design tool in human factors research.

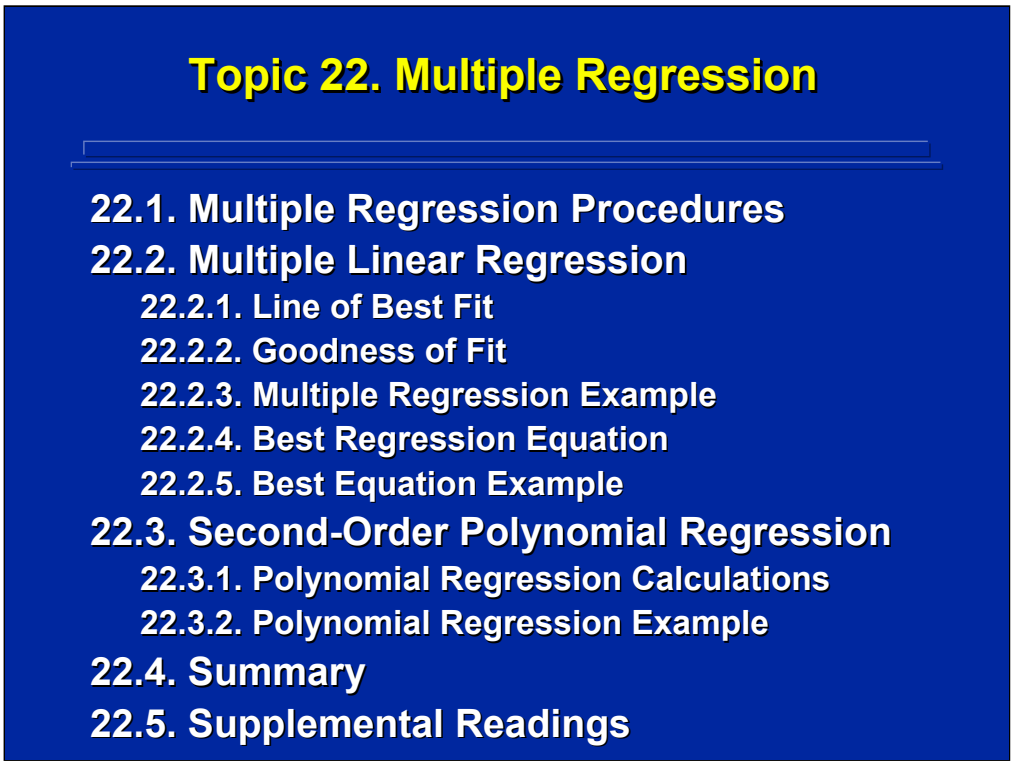



The focus of ANOVA designs is statistical hypothesis testing. All the topics in Section 5 use experiments to collect data for building empirical models that predict human performance in complex systems. These predictions can be used for interface design tradeoffs. The data from these experiments can also be combined into integrated databases describing complex systems.

The next three topics describe the details of techniques used in empirical model building. Topic 22 describes polynomial regression which is a general form of multiple regression used to specify the functional relationship of the empirical model. Topic 23 describes a useful second-order experimental design to collect the necessary and sufficient data for generating empirical models. And, finally, Topic 24 describes the concept of sequential experimentation in which an integrated set of small experiments are used to build empirical models that include a large number of factors. The results of these sequential experiments form an integrated database of research rather that a series of isolated experiments that cannot be related to each other.




The two chapters by Box, Hunter, and Hunter (1978) provide the classic distinction between empirical and mechanistic models and the use of experimental designs and sequential experimentation in building quantitative models. Both the Box and Draper (1987) and the Box, Hunter, and Hunter (2005) texts listed on this slide provide a general overview of quantitative models in the form of mechanistic and empirical models. Chapter 1 of both Box and Draper (1987) and Myers and Montgomery (2002) introduce the concept of empirical model building through experimentation. The chapters by Wickens (1992) provide details on normative models used in human factors and ergonomics.




Topic 22 provides an overview of multiple regression procedures used to generate empirical models that involve more than one factor. First, multiple linear regression is described to demonstrate the calculations involved when considering more than one factor in the empirical model. Next, second-order polynomial regression is discussed as the general form for stating empirical models in human factors research. A summary of these procedures as well as additional readings for details on multiple regression are provided at the end of this topic.

## 22.1. Multiple Regression Procedures

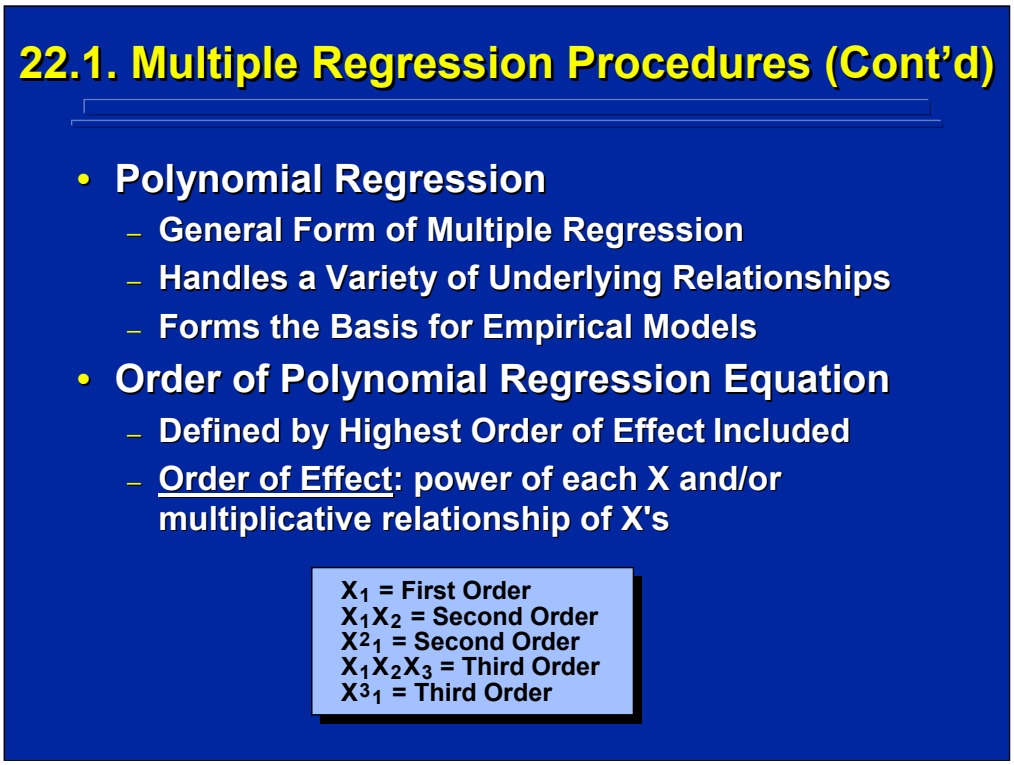
- **Extension of Simple Regression**
  - Multiple Predictors, $X_i$
  - Partial Regression Beta Weight, $b_i$, for Each Predictor
- **Multiple Regression Calculations**
  - Line of Best Fit
  - Goodness of Fit
  - Best Equation
- **Polynomial Function**
  - Two Predictor Example

    $Y' = b_0 + b_1X_1 + b_2X_2 + b_3X^2_1 + b_4X^2_2 + b_5X_1X_2$
  - Linear Beta Weights

Multiple regression is an extension of simple regression discussed in Topic 19 by including more than one predictor, $X_i$, in the regression equation. Each predictor has its own beta weight called a partial regression weight in multiple regression.

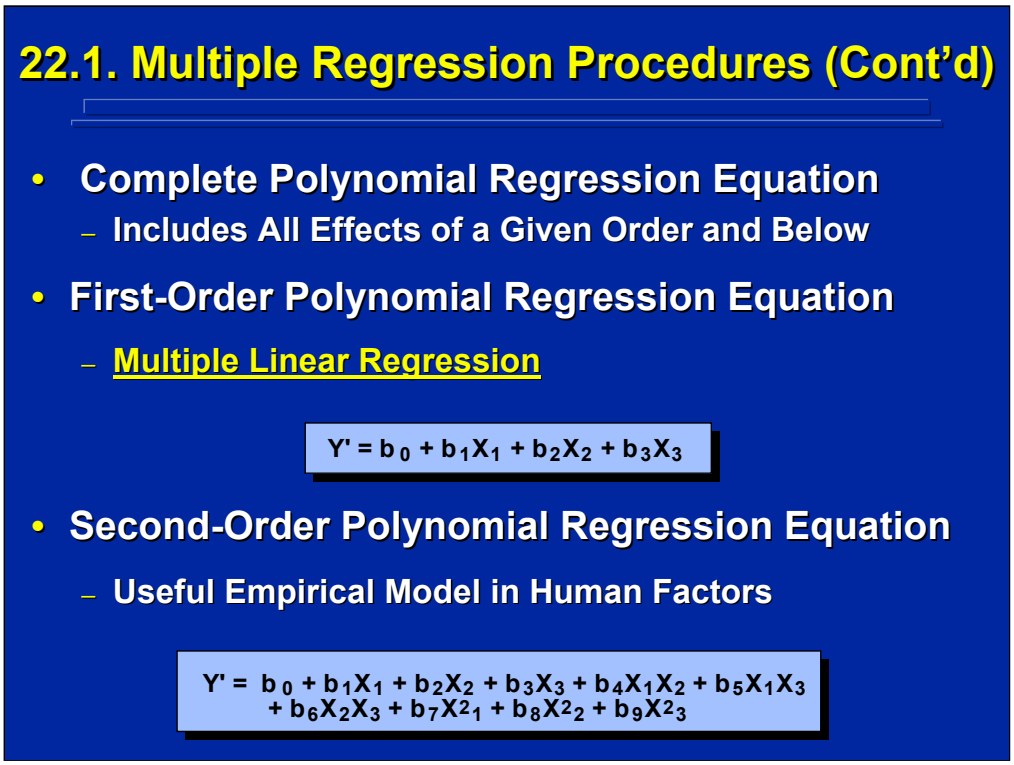
There are three major areas of multiple regression calculations. The line of best fit involves the least square calculations on the partial regression weights in the multiple regression. The goodness of fit determines how well the multiple regression represents the data. And, the best equation determines the optimal number of predictors to include in the multiple regression.

Polynomial functions allow each $X_i$ to represent more than linear effects. The regression equation shown at the bottom of this slide is an example of a polynomial function with two X's. Note that the polynomial regression includes the linear effects, the cross-product effect, and the quadratic effects of $X_1$ and $X_2$. The beta weights, $b_i$, are all linear weights in this polynomial regression equation.




Polynomial regression is the general form of multiple regression that includes linear and non-linear effects of predictors in the regression equation. These regression equations can be used to describe a variety of underlying relationships affecting operator performance in complex systems. Consequently, polynomial regression forms the basis for stating empirical models that predict human performance.

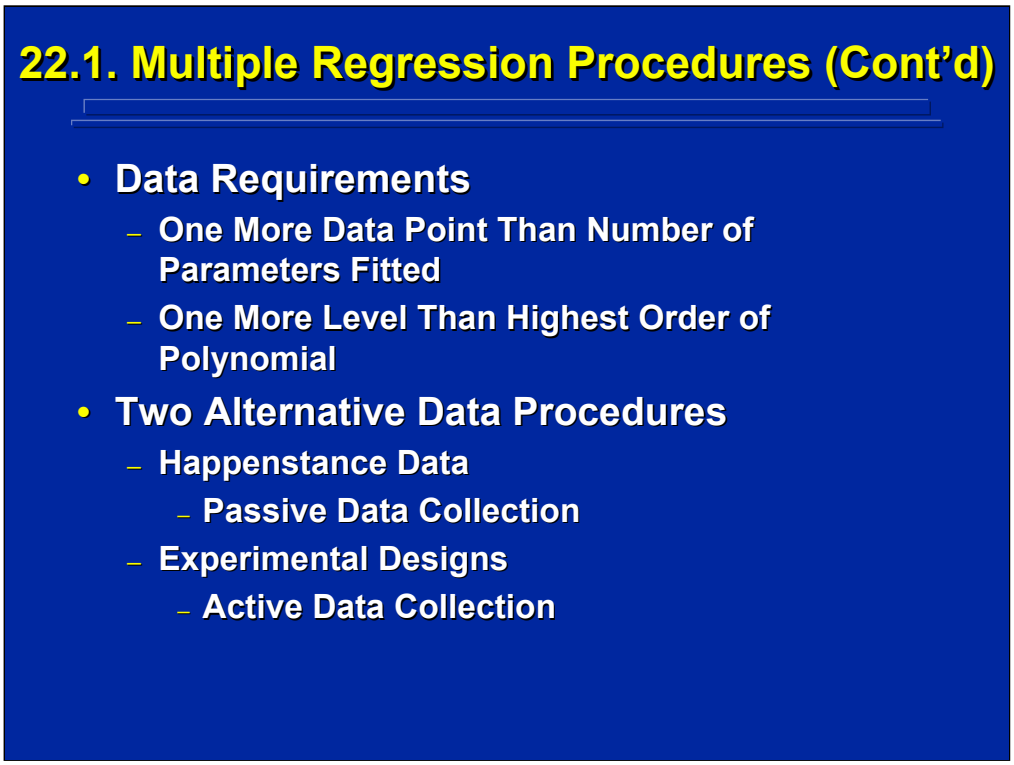
A polynomial regression function is defined by the highest order used in the equation. Order is determined by the power of each predictor, X, and/or the multiplicative relationship of the Xs. Several examples of first-, second-, and third-order effects are shown on the bottom of this slide.




If all the highest-order effects and all the lower-order effects of a set of Xs are included in the polynomial, it is called complete. If not, the polynomial is called incomplete. For example, the first-order and second-order polynomials for three Xs shown on this slide are both complete.

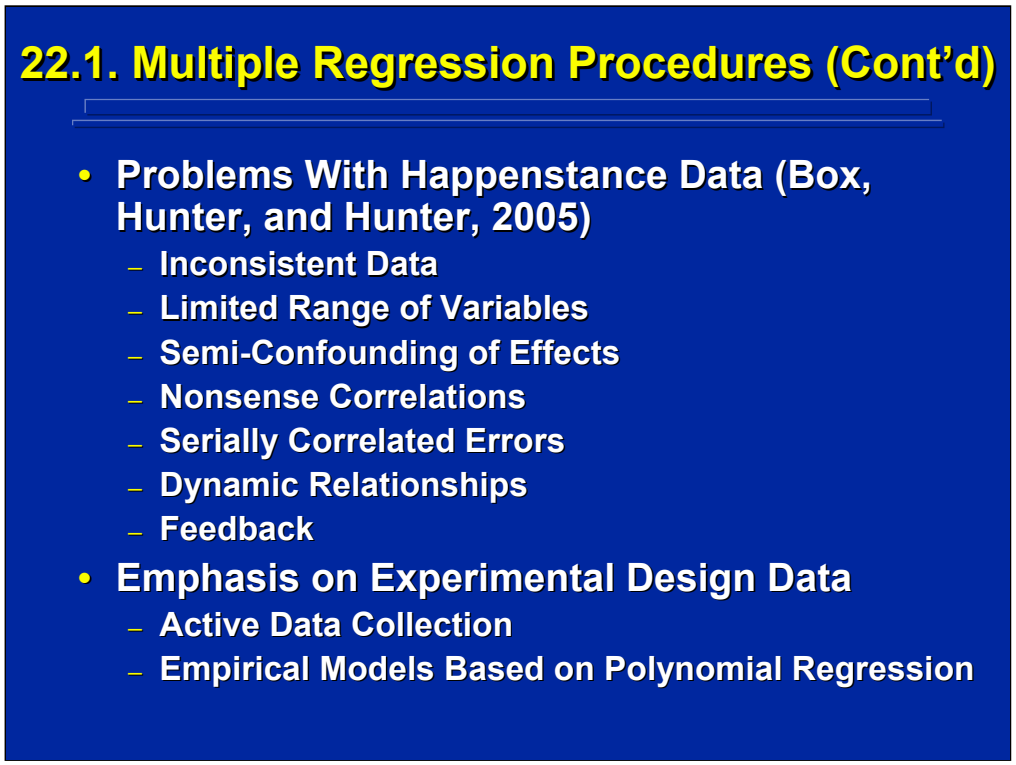
There are two general types of multiple regression equations used in human factors. Multiple linear regression is actually a sub-set of polynomial regression in which all the predictors form a first-order polynomial as shown in the middle portion of this slide for three Xs. Multiple linear regression is the most common form of multiple regression used in behavioral research.

Higher-order polynomials can represent curvilinear regression as shown by the second-order polynomial at the bottom of this slide for three Xs. The second-order effects of the linear-by-linear interaction effects (i.e., $X_iX_j$) weighted by the partial regression weights $b_4$, $b_5$, and $b_6$ plot sloping planes; whereas, the pure quadratic effects weighted by the partial regression weights $b_7$, $b_8$, and $b_9$ plot quadratic effects in the polynomial regression. Since two-way interactions which include linear-by-linear effects are important in human factors research, second-order empirical models are often used.

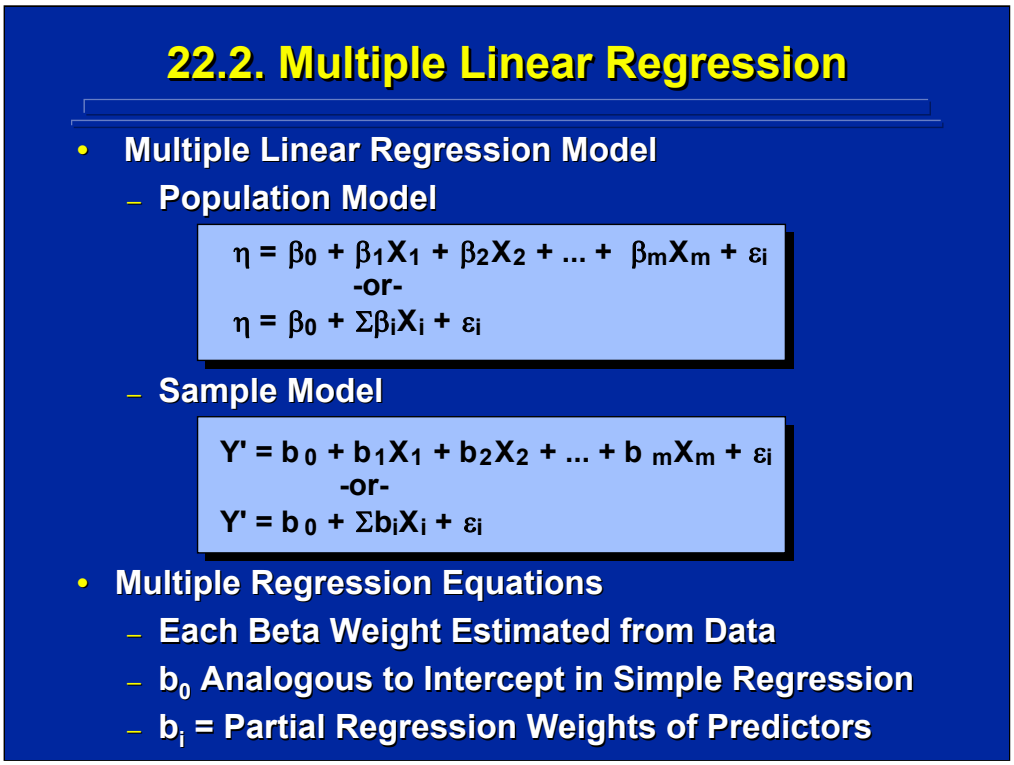



Two minimum data requirements need to be considered in conducting multiple regression. First, one more data point is needed than the number of parameters in the equation. Second, one higher level needs to be observed than the highest-order effect of a predictor. For example, in the three factor, second-order polynomial regression shown on the previous slide, a minimum of eleven different data points involving three levels of each of the two factors is needed to determine the ten beta weights in the empirical model.

Data for multiple regression analysis can be collected by happenstance or through experimental designs. Happenstance data are obtained through passive data collection from either data archives that already exist or data that is observed when the levels of factors are not controlled. Alternatively, experimental designs can be used to control the levels of the factors in an active data collection procedure.

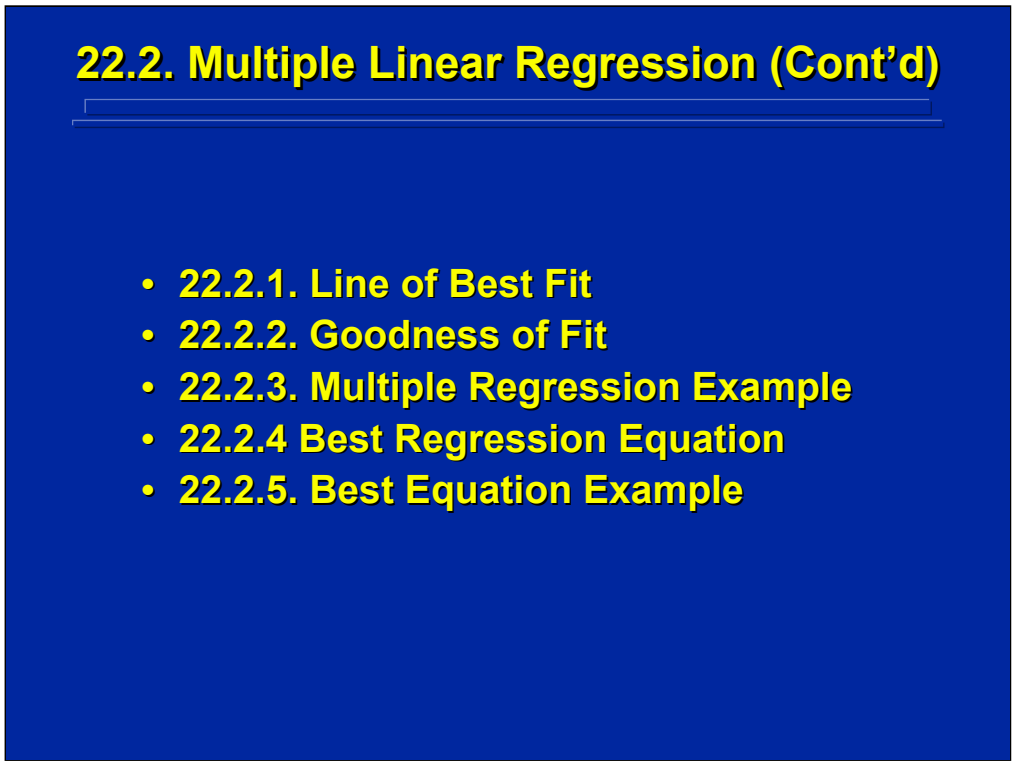



Box, Hunter, and Hunter (2005, pp. 397-406) point out seven potential problems as shown on this slide that can occur in using happenstance data for generating empirical models based on the multiple regression of several variables using polynomial regression. All of these problems can be either avoided, or controlled, by using experimental designs to collect data for empirical model building. Consequently, special purpose experimental designs and sequential experimentation for building empirical models are stressed in Section 5.




Multiple linear regression is a subset of multiple regression in which only the linear effect of each of several factors are included in the regression model. The population and sample models of multiple linear regression are shown on the top portion of this slide. Sample data are used to determine the beta weights in the sample model which, in turn, provide the best estimates of the population regression model.

Calculation of the beta weights is the major computational procedure in solving the multiple regression. The $b_0$ value is analogous to the intercept value in simple regression as described in Topic 19. The $b_i$ values are called partial regression weights and represent the empirically determined weights for each of the factors or predictors considered in the multiple linear regression model.




This subsection describes the least squares solution for the line of best fit, metrics for assessing the goodness of fit, and several techniques for choosing the best subset of predictors in the multiple linear regression equation. An example using happenstance data is provided to demonstrate the use of these procedures in conducting a multiple linear regression and choosing the best regression equation.

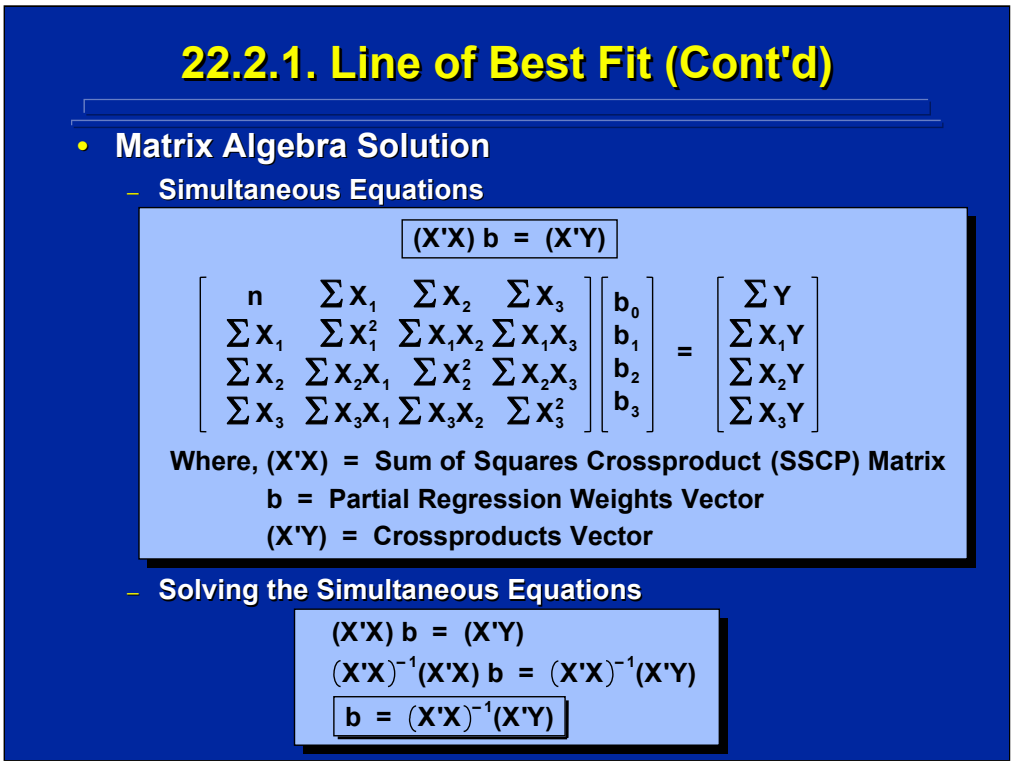
## 22.2.1. Line of Best Fit

- **Least Squares Criterion, Q**
  - $Q = (Y - Y')^2$ is a Minimum
- **Three Predictor Example**
  - Equation: $Y' = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e_i$
  - Minimize Q
    - $Q = [Y - (b_0 + b_1X_1 + b_2X_2 + b_3X_3)]^2$
- **"Normal" Equations**
  - Set Partial Derivatives to 0
  - Vector of Residuals is Normal to Vector of "X" Variables
  - Simultaneous Equations for Three Predictors

$$\begin{cases} nb_0 + (\Sigma X_1)b_1 + (\Sigma X_2)b_2 + (\Sigma X_3)b_3 = \Sigma Y \\ (\Sigma X_1)b_0 + (\Sigma X_1^2)b_1 + (\Sigma X_1X_2)b_2 + (\Sigma X_1X_3)b_3 = \Sigma X_1Y \\ (\Sigma X_2)b_0 + (\Sigma X_2X_1)b_1 + (\Sigma X_2^2)b_2 + (\Sigma X_2X_3)b_3 = \Sigma X_2Y \\ (\Sigma X_3)b_0 + (\Sigma X_3X_1)b_1 + (\Sigma X_3X_2)b_2 + (\Sigma X_3^2)b_3 = \Sigma X_3Y \end{cases}$$

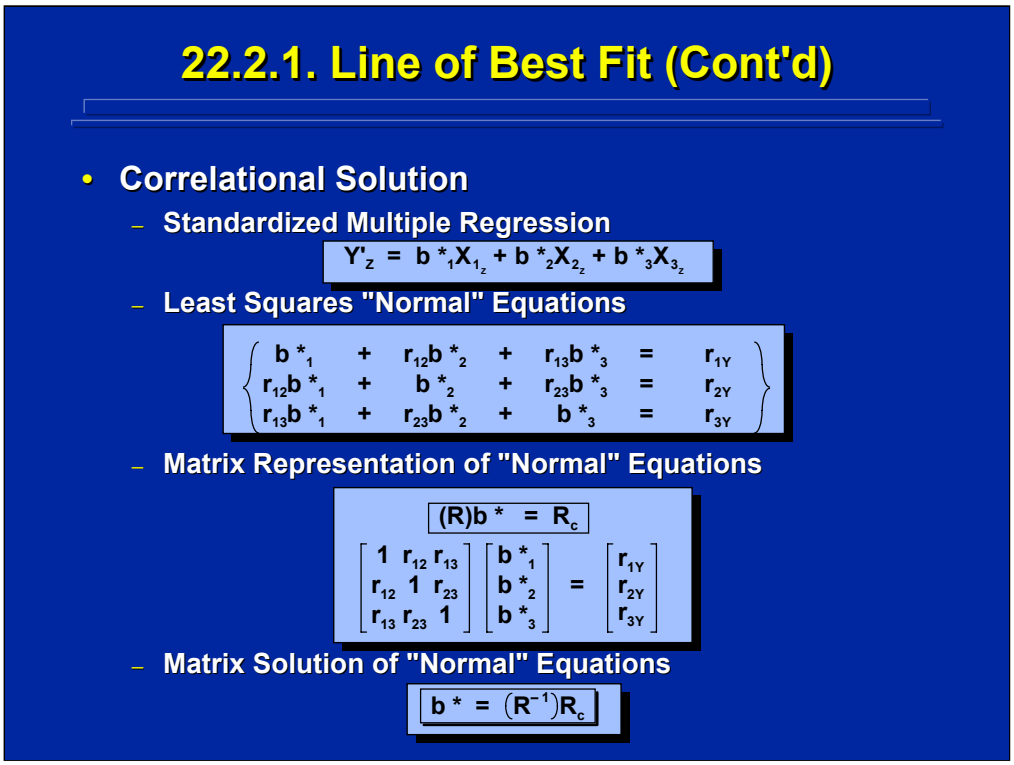
The least squared criterion is used to determine the line of best fit in multiple regression. This criterion minimizes the sum of squared differences between the observed and predicted values of Y. An example of a multiple linear regression for Y' with $b_0$ and three predictors is shown in the middle of this slide. A least squares solution requires taking the difference between Y and the partial derivatives of the prediction equation with respect to each of the four unknowns and setting them equal to zero.

The resulting partial derivatives yield four simultaneous equations that can be solved to determine the values of the four parameters, $b_0$ to $b_3$ in a multiple regression with three predictors. These simultaneous equations are called “normal” equations because the vector of residuals is orthogonal or perpendicular (i.e., normal) to the vector of “X” variables according to the least squares criterion, Q.



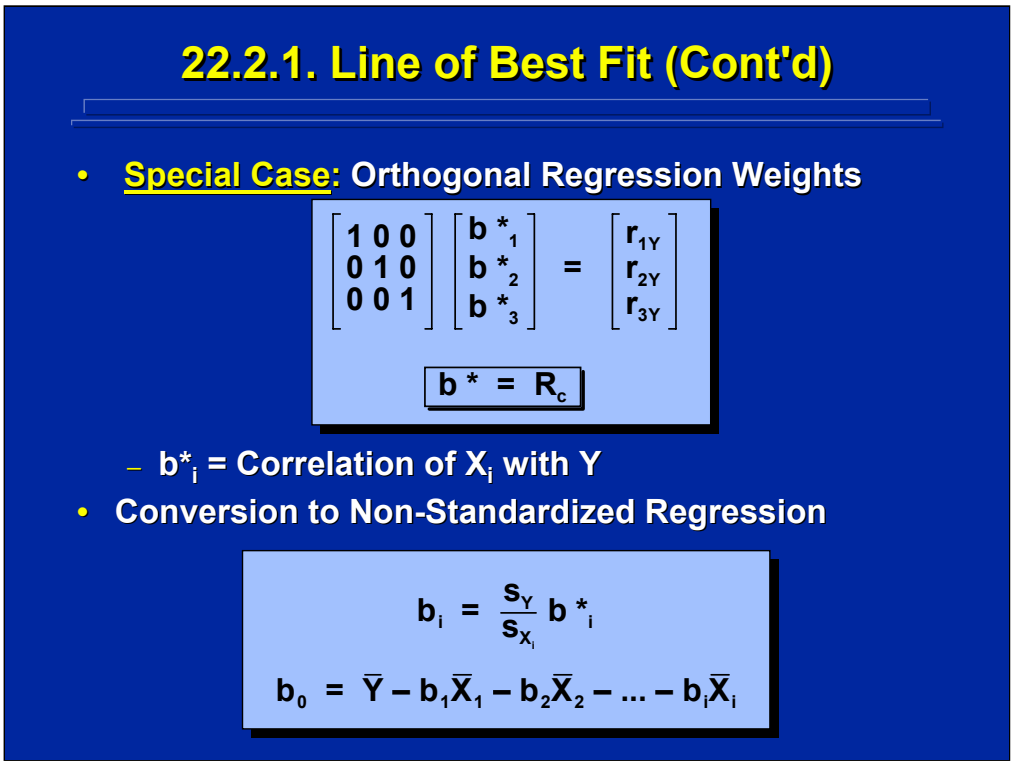

Matrix algebra is used to solve the normal equations in multiple regression. See Draper and Smith (1981, Chapter 2), Myers (1990, Appendix A), Winer, Brown, & Michels (1991, Appendix B) for a review of matrix algebra procedures used in regression. This slide shows the matrix algebra representation of the four simultaneous equations of normal equations presented on the previous slide. Note this set of simultaneous equations is simply represented in matrix notation by the product of the sum of squares crossproducts (SSCP) matrix, (X'X), and the b vector equals the crossproducts vector, X'Y.

The matrix algebra solution for these simultaneous equations is simply to multiply the X'X matrix by its inverse, $(X'X)^{-1}$. As shown on the bottom portion of this slide, the beta weights equal the inverse of the X'X matrix times the crossproducts vector. Inverting the SSCP matrix can become tedious, and multiple regression solutions are usually conducted through computer analysis.



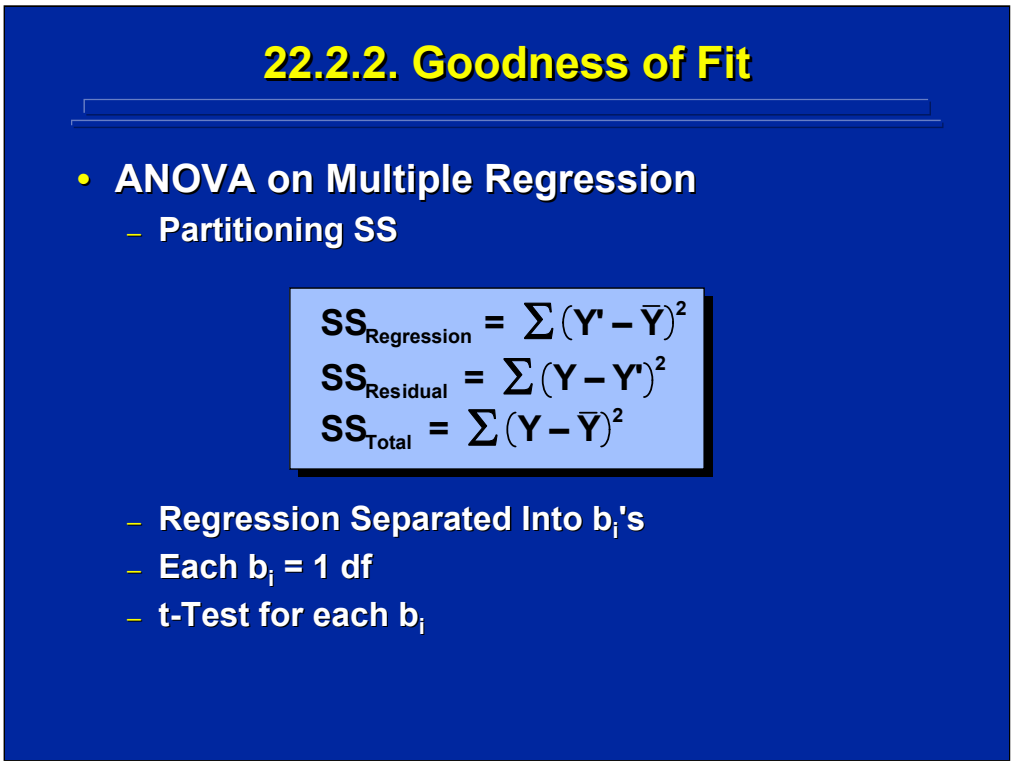

This slide summarizes the correlational solution for multiple regression in the special case were the $X_i$ predictors are standardized. Note that in standardized multiple regression the predicted value is also the standardized, or Z-score of Y, not the Y score. For standardized scores, the $b_0$ value is equal to 0 and the other beta weights are designated as $b_i$*. The normal equations and matrix algebra representations of these simultaneous equations can be stated in terms of correlations as shown in the middle portions of this slide.

As shown on the bottom of this slide, the matrix algebra solution for the standardized regression weights is simply the intercorrelation matrix, R, pre-multiplied by the inverse of the intercorrelation matrix, $R^{-1}$.



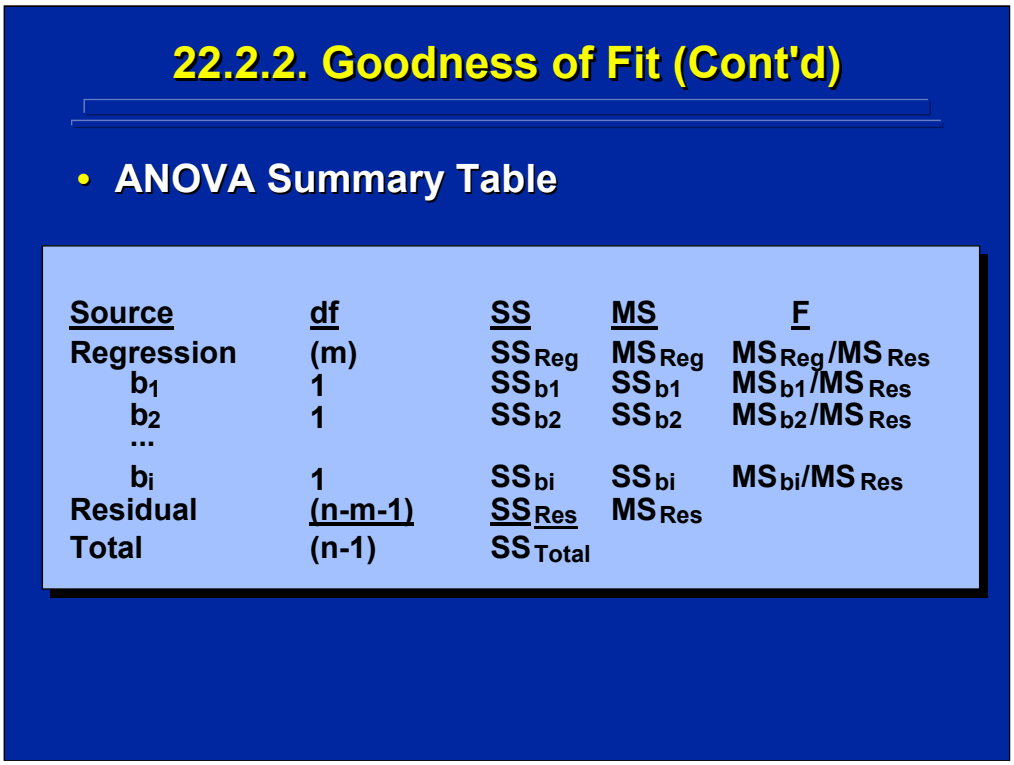

A special case of standardized regression occurs when the multiple predictors are independent of each other. Their intercorrelation equal 0 and the intercorrelation matrix becomes diagonal. As shown on the top portion of this slide, the standardized beta weights, $b^*_i$, simply equal the correlation of each predictor with Y.

The lower portion of this slide shows the conversion of non-standardized to standardized regression weights as well as the formula for determining $b_0$ in non-standardized multiple regression. Both the standard deviations of the Y scores and the various $X_i$ scores are needed to make this conversion.

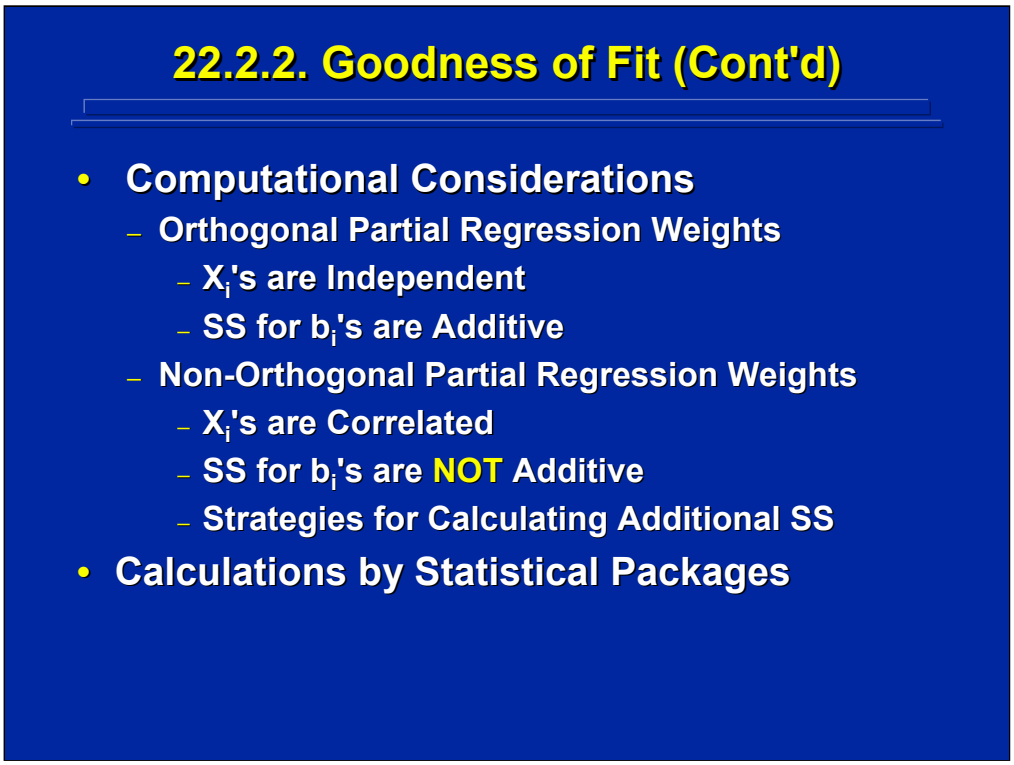



Several procedures exist for determining the goodness of fit of multiple linear regression equations. An ANOVA of the regression equation can be conducted to test the overall significance of the regression model and the individual partial regression weights can be tested for significance. Similar to simple regression, the total sum of squares is divided into two additive parts, regression and residual. The overall regression can then be tested by residual as the error term.

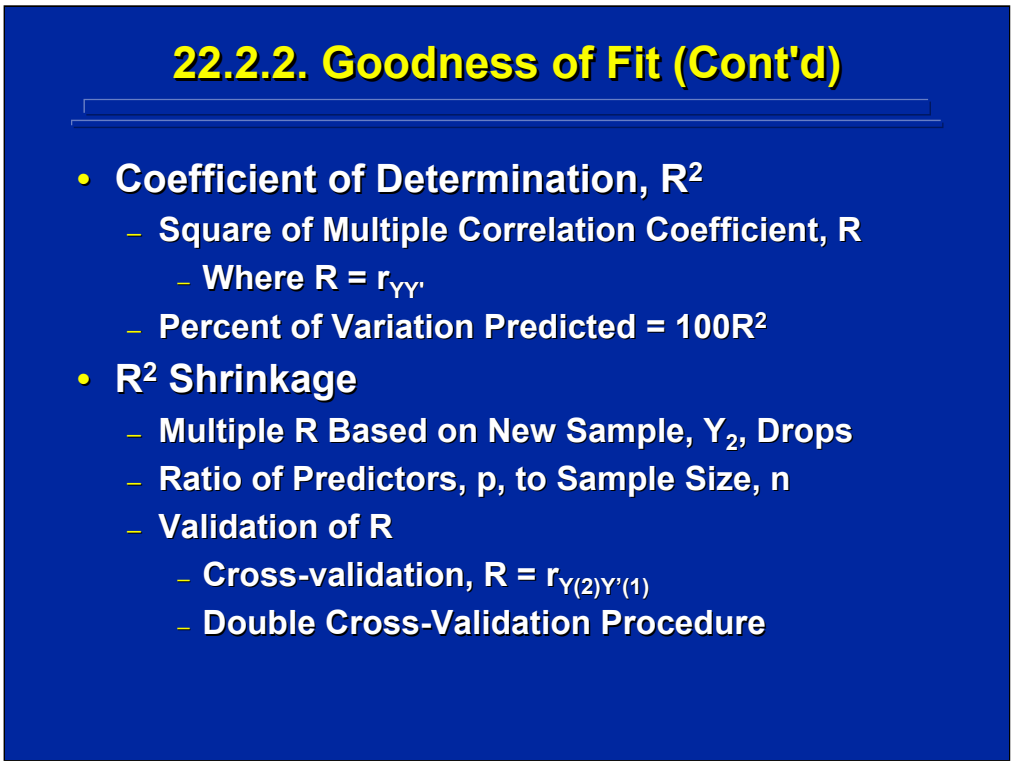
Regression can also be separated into the effects of individual partial regression weights, $b_i$, to determine if any of the various predictors account for a significant amount of variation. Since each partial regression weight has one degree of freedom, a simple t-test can be used to test each $b_i$.




This slide provides a general ANOVA Summary Table layout for conducting an ANOVA on multiple linear regression. An overall regression model of “m” predictors can be subdivided into the partial regression weights considered in the overall model. If the partial regression weights are either independent or are tested considering all the other beta weights are present, the $MS_{Residual}$ can be used as the error term to test the significance each beta weight as well as the overall regression model.




Intercorrelation of the predictors is a major computational consideration in testing the significance of the partial regression weights. If the predictors are independent, the sum of squares for the partial regression weights are additive and will equal the sum of squares regression. This would occur if the data used to generate the multiple linear regression were drawn from a $2^k$ factorial design. If the partial regression weights are non-orthogonal, however, their partial sums of squares are not additive. Correlations among the predictors can greatly affect the partial regression weights when the regression model is based on happenstance data from observational studies where levels of the factors are not controlled during data collection. Alternative strategies for testing partial regression weights need to be considered in conducting t-tests and F-tests on the beta weights in order to consider the partial correlation effects. Consequently, statistical analysis packages are usually used to conduct goodness of fit significance tests in multiple linear regression.




Goodness of fit can be assessed by the multiple correlation coefficient, R, which is the correlation of the observed score with the predicted score of the multiple regression equation. The coefficient of determination is equal to $R^2$ and is the percent of variation predicted by regression when multiplied by 100.

The percent of variation predicted by the original multiple regression model is usually expected to drop (shrink) when the regression model is extended to a new data set (Pedhazur 1982, pp. 147-150) due to unique characteristics in small samples. The extent of shrinkage is a function of the number of predictors, p, and the sample size, n, used to develop the model. In general, shrinkage increases as the p/n ratio increases.

Validation procedures can be used to choose the best subset of predictors in regression in order to reduce shrinkage and increase the validity of the regression equation. In cross-validation the observed values of a new sample of data are correlated with the predicted regression model values from the original sample. In double cross-validation the predicted value based on one sample is correlated with the observed value from a second sample and vice versa.

## 22.2.2. Goodness of Fit (Cont'd)

- **Adjusted Coefficient of Determination, $R_{adj}^2$**
  - Percent of Variation Expected with Shrinkage
  - Estimate of Shrinkage (Pedhauzer, 1982)

    $R_{adj}^2 = 1 - [1 - R^2][(n - 1)/(n - p - 1)$

    where, n = sample size

    p = number of parameters including $b_0$
  - Estimate of Shrinkage (SAS, 2004)

    $R_{adj}^2 = 1 - [1 - R^2][(n - i)/(n - p)$

    where, i = 1 if model includes intercept; if not i = 0
  - Minor Difference when n > 50

Rather than conducting cross-validation studies, an adjusted Coefficient of Determination, $R_{adj}^2$, can be used to estimate regression shrinkage. This estimate is based on the number of predictors, p, in the regression model, and the sample size, n, used to generate the multiple regression equation. The formulae for two such estimates are shown on this slide. The difference in estimates is quite small among these formulae when sample size is greater than 50.

## 22.2.3. Multiple Regression Example

- **Example Problem: The commander's combat operation performance in a battalion level command and control center for the Army is scored on a 100 point scale. Scores of fifteen battalion commanders are predicted as a function of four command and control tasks. The predictors are the time to complete Recognition, Decision, Communication, and, Evaluation tasks. What is the linear relationship of these four tasks on predicting the performance score? Are any of these predictors significant ($p < 0.05$)?**

(Click in this red rectangle to see SAS calculations for this example.)

The example problem described on this slide uses happenstance data based on an observational study. The various levels of the four command and control tasks as measured by time to complete each subtask are merely observed, not controlled. The observational data of 15 commanders are then used to calculate a multiple linear regression model predicting the commanders' performance scores on a 100 point scale.

## 22.2.3. Multiple Regression Example (Cont'd)

- **Example Problem Data Set**

| Recognition Task (Rec) | Decision Task (Dec) | Communication Task (Com) | Evaluation Task (Eval) | Performance Score (PS) |
|---|---|---|---|---|
| 56 | 47 | 59 | 55 | 76 |
| 60 | 49 | 57 | 53 | 80 |
| 59 | 50 | 64 | 57 | 86 |
| 52 | 55 | 52 | 54 | 75 |
| 51 | 45 | 55 | 58 | 66 |
| 54 | 58 | 53 | 60 | 76 |
| 60 | 49 | 57 | 62 | 90 |
| 57 | 50 | 54 | 53 | 71 |
| 58 | 53 | 56 | 54 | 77 |
| 53 | 57 | 53 | 56 | 79 |
| 63 | 45 | 54 | 51 | 83 |
| 54 | 53 | 55 | 50 | 70 |
| 58 | 50 | 58 | 55 | 76 |
| 60 | 50 | 57 | 55 | 75 |
| 55 | 56 | 50 | 53 | 73 |

(Click in this red rectangle to see SAS calculations for this example.)

This slide presents the hypothetical data of the time each of the 15 commanders took to finish the Recognition, Decision, Communication, and Evaluation subtasks as well as their overall performance score on the command and control exercise.

## 22.2.3. Multiple Regression Example (Cont'd)

- **Least Squares "Normal" Equations**

$$\begin{cases} nb_0 + (\Sigma X_1)b_1 + (\Sigma X_2)b_2 + (\Sigma X_3)b_3 + (\Sigma X_4)b_4 = \Sigma Y \\ (\Sigma X_1)b_0 + (\Sigma X_1^2)b_1 + (\Sigma X_1X_2)b_2 + (\Sigma X_1X_3)b_3 + (\Sigma X_1X_4)b_4 = \Sigma X_1Y \\ (\Sigma X_2)b_0 + (\Sigma X_2X_1)b_1 + (\Sigma X_2^2)b_2 + (\Sigma X_2X_3)b_3 + (\Sigma X_2X_4)b_4 = \Sigma X_2Y \\ (\Sigma X_3)b_0 + (\Sigma X_3X_1)b_1 + (\Sigma X_3X_2)b_2 + (\Sigma X_3^2)b_3 + (\Sigma X_3X_4)b_4 = \Sigma X_3Y \\ (\Sigma X_4)b_0 + (\Sigma X_4X_1)b_1 + (\Sigma X_4X_2)b_2 + (\Sigma X_4X_3)b_3 + (\Sigma X_4^2)b_4 = \Sigma X_4Y \end{cases}$$

- **Matrix Representation of "Normal" Equations**

$$[X'X][b] = [X'Y]$$

$$\begin{bmatrix} 15 & 850 & 767 & 834 & 826 \\ 850 & 48334 & 43371 & 47332 & 46788 \\ 767 & 43371 & 39453 & 42552 & 42247 \\ 834 & 47332 & 42552 & 46528 & 45959 \\ 826 & 46788 & 42247 & 45959 & 45628 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} = \begin{bmatrix} 1153 \\ 65532 \\ 58922 \\ 64234 \\ 63592 \end{bmatrix}$$

(Click in this red rectangle to see SAS calculations for this example.)

The top of this slide shows the normal equations for determining the least squares solution to the multiple linear regression equation with four predictors and an intercept value. The bottom portion of this slide shows these normal equations in matrix notation using the data from the previous slide to calculate the X'X matrix and the X'Y vector using SAS (2004) as described in the Slater and Williges (2006) appendix.

## 22.2.3. Multiple Regression Example (Cont'd)

- **Matrix Solution of Simultaneous Equations**

$$[b] = [X'X]^{-1}[X'Y]$$

$$\begin{bmatrix} b_0 = -85.83 \\ b_1 = 1.40 \\ b_2 = 0.48 \\ b_3 = 0.29 \\ b_4 = 0.78 \end{bmatrix} = \begin{bmatrix} 97.0109 & -0.5213 & -0.5603 & -0.3776 & -0.3225 \\ -0.5213 & 0.0086 & 0.0021 & -0.0031 & 0.0017 \\ -0.5603 & 0.0021 & 0.0063 & 0.0030 & -0.0009 \\ -0.3776 & -0.0031 & 0.0030 & 0.0101 & -0.0030 \\ -0.3225 & 0.0017 & -0.0009 & -0.0030 & 0.0080 \end{bmatrix} \begin{bmatrix} 1153 \\ 65532 \\ 58922 \\ 64234 \\ 63592 \end{bmatrix}$$

- **Multiple Linear Regression Equation**

**PS = - 85.83 + 1.40Rec + 0.48Dec + 0.29Com + 0.78Eval**

- **Coefficient of Determination**
  - $R^2 = 0.70$

(Click in this red rectangle to see SAS calculations for this example.)

The matrix solution for the multiple linear regression equation based on the example problem data is presented on the top of this slide and was calculated by SAS (2004) and described in the Slater and Williges (2006) appendix. The resulting multiple linear regression model predicting overall performance score (PS) as a function of completion times of the four predictor subtasks is shown in the center of this slides. This equation accounts for 70% of the performance variation as determined by the Coefficient of Determination shown at the bottom of this slide.

## 22.2.3. Multiple Regression Example (Cont'd)

- ANOVA on Multiple Regression

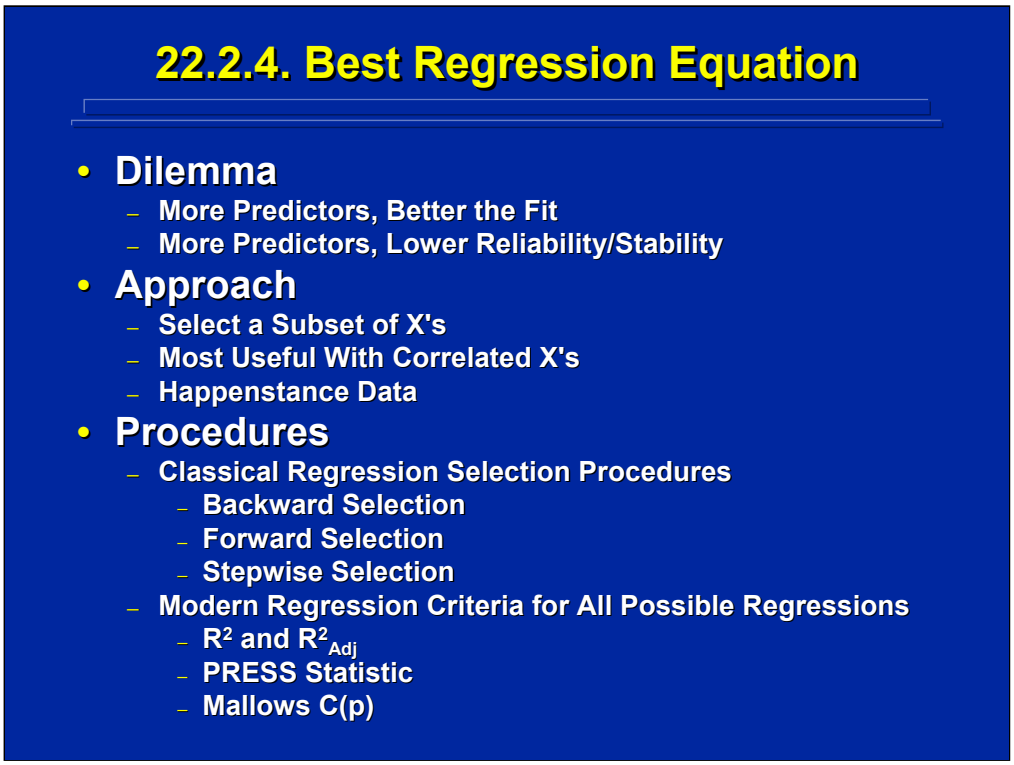
| Source | df | SS | MS | F |
|---|---|---|---|---|
| Regression | (4) | 370.64 | 92.66 | 5.75* |
| $b_{Rec}$ | 1 | 225.21 | 225.21 | 13.98* |
| $b_{Dec}$ | 1 | 36.62 | 36.62 | 2.27 |
| $b_{Com}$ | 1 | 8.28 | 8.28 | 0.51 |
| $b_{Eval}$ | 1 | 75.91 | 75.91 | 4.71 |
| Residual | 10 | 161.09 | 16.11 | |
| Total | 14 | 531.73 | | |

*$p < 0.05$

(Click in this red rectangle to see SAS calculations for this example.)

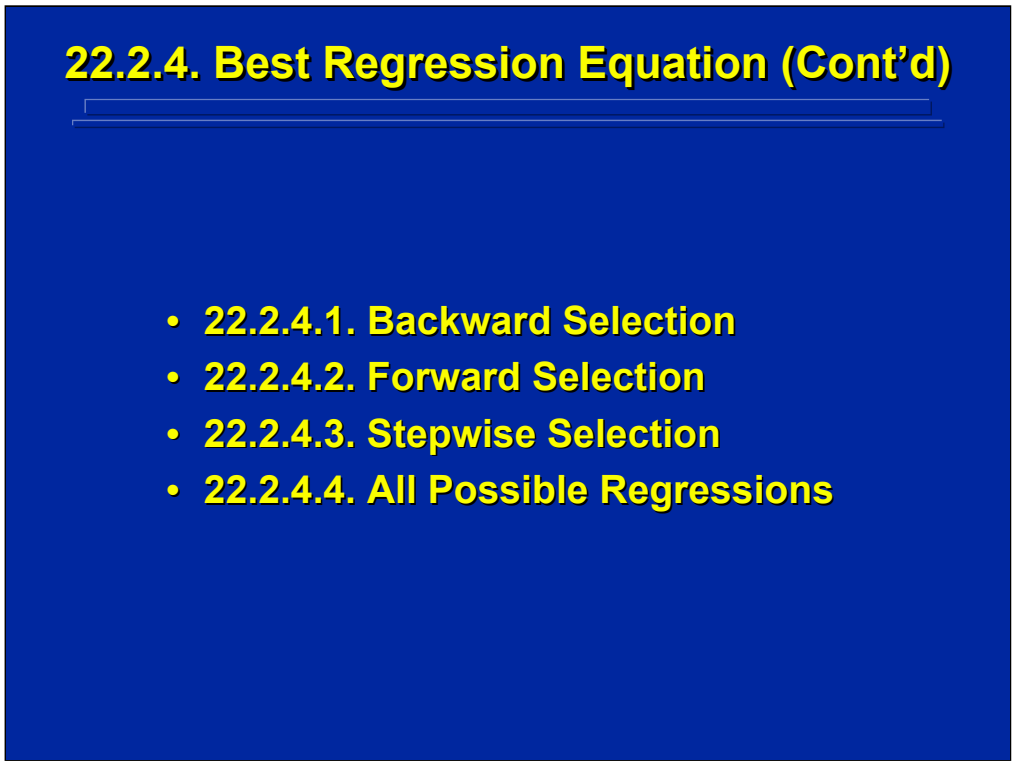
The Summary Table for the ANOVA on regression is shown on this slide. The overall multiple linear regression model with four predictors is significant at the 0.05 level. Of the four predictors, however, only the target recognition subtask is a significant ($p < 0.05$) predictor of a commander's command and control performance score. Note that the total of the sum of squares of the four partial regression weights is 346.02 and does not equal the sum of squares of regression (370.64) due to the covariance among the four predictors resulting from the happenstance data.

The F-tests on these partial regression weights is based on Type III SS calculated by SAS (2004) to test the unique contribution of each beta weight given that the other three beta weights exist in the model. These F-tests provide the same results as the square of the t-tests of significance on each partial regression weight as described in the Slater and Williges (2006) appendix.




A linear regression model that includes all predictors investigated may not be the best model in terms of reliability and validity. Choosing the appropriate number of predictors to use in a final multiple linear regression model is a complicated task. As more predictors are added to a multiple regression, the percent of variation predicted increases. At the same time, the shrinkage increases and validity of the multiple regression decreases as the number of predictors increase. When happenstance data are used, the parameters are often correlated and the covariance among predictors also needs to be considered in choosing the best regression equation. Consequently, the experiment must choose the best subset of parameters to use in the multiple linear regression equation.

In this sub-section, several regression procedures are presented that can be used to choose the best multiple linear regression equation. Classical regression techniques refer to backward, forward, and stepwise selection procedures. Modern regression procedures consider all possible regression equations and evaluate them in terms of tradeoffs using statistics such as the Adjusted Coefficient of Determination, $R^2_{Adj}$, the PRESS statistic, and Mallows C(p) value.



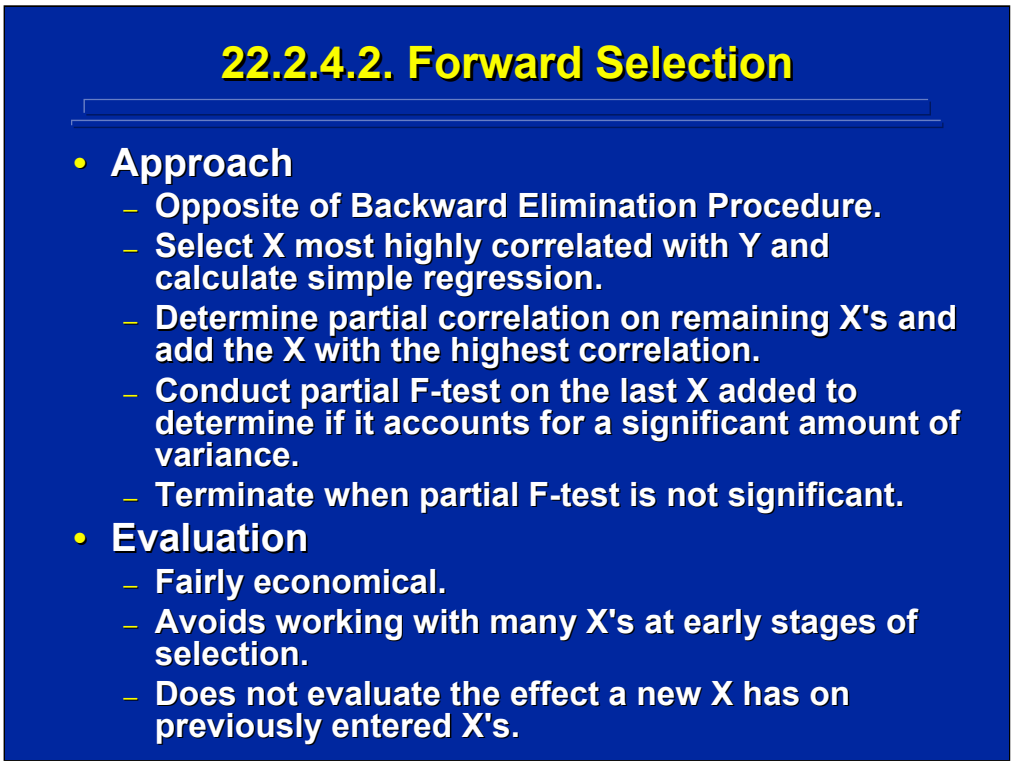

The backward, forward, and stepwise selection procedures and modern regression statistics for evaluating all possible regression equations are described in the following slides as alternatives for choosing the best possible multiple linear regression equation.

## 22.2.4.1. Backward Selection

- **Approach**
  - Calculate regression with all X's
  - Conduct partial F-test on each variable assuming it was last entered.
  - Compare lowest F-ratio, FL, to a pre-selected level of significance, $F_0$. Eliminate X if $F_L < F_0$.
  - Repeat steps 1-3 until $F_L > F_0$. Accept regression equation at this point.
- **Evaluation**
  - Good procedure if interested in seeing regression equation with all predictors.
  - For a near singular X'X matrix, rounding errors can give nonsense results.

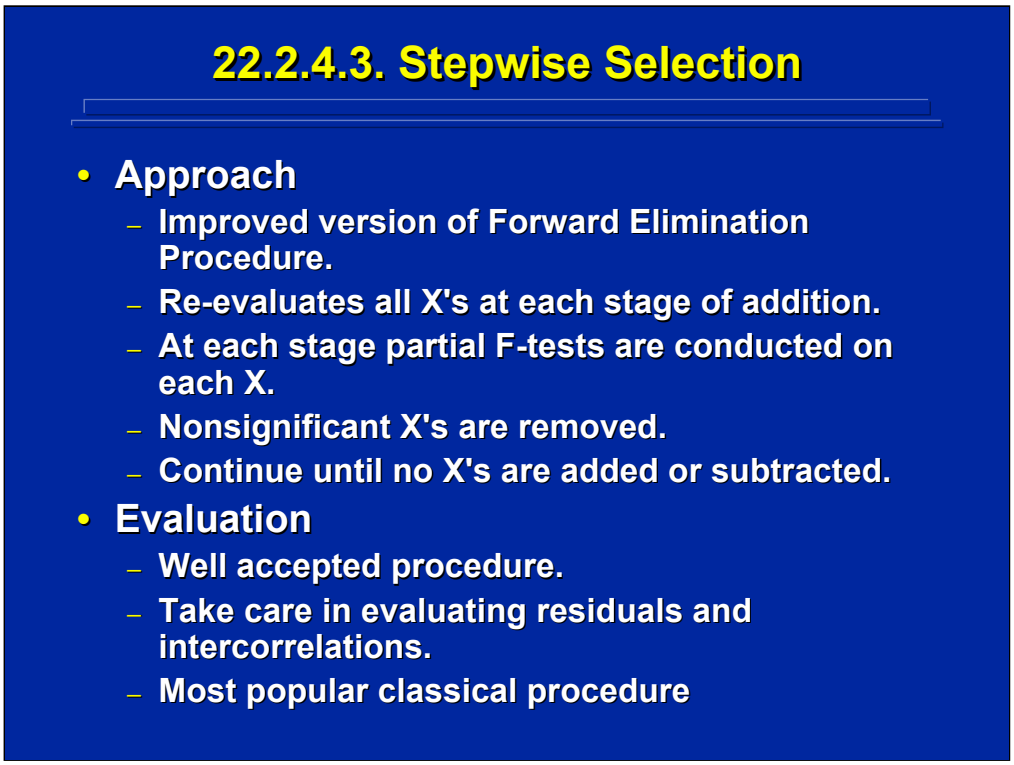
The backward selection procedure begins with a multiple linear regression that includes all possible predictors and then conducts partial F-tests to remove one predictor at a time until only significant predictors remain in the regression equation. The calculations used with this procedure are summarized on the slide.

Backward selection is a useful technique if the experimenter wants to begin with a regression equation that includes all of the predictors. When two predictors have close to zero correlation, however, rounding errors can cause spurious results.



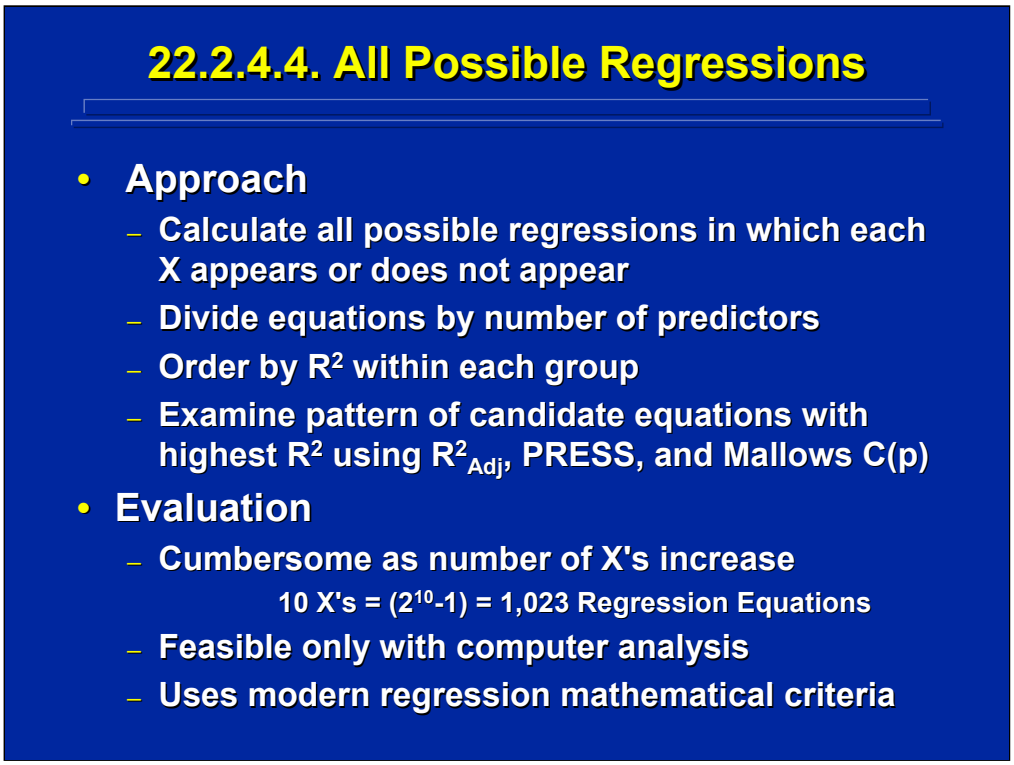

The forward selection procedure is the opposite of backward selection in that it begins with only the highest predictor correlated with performance in the regression equation and then adds one additional predictor at a time to the multiple linear regression equation. The specific procedural steps and partial F-tests associated with forward selection are summarized on this slide.

Forward selection is economical because it begins with a simple regression equation and then progresses to more complex multiple linear regression equations. When a new predictor is added, however, the effect of this predictor effect on previously entered predictors is not evaluated. Depending on the covariance among predictors, some previously added predictors may no longer contribute significantly to the multiple linear regression equation when used in combination with the newly added predictors.



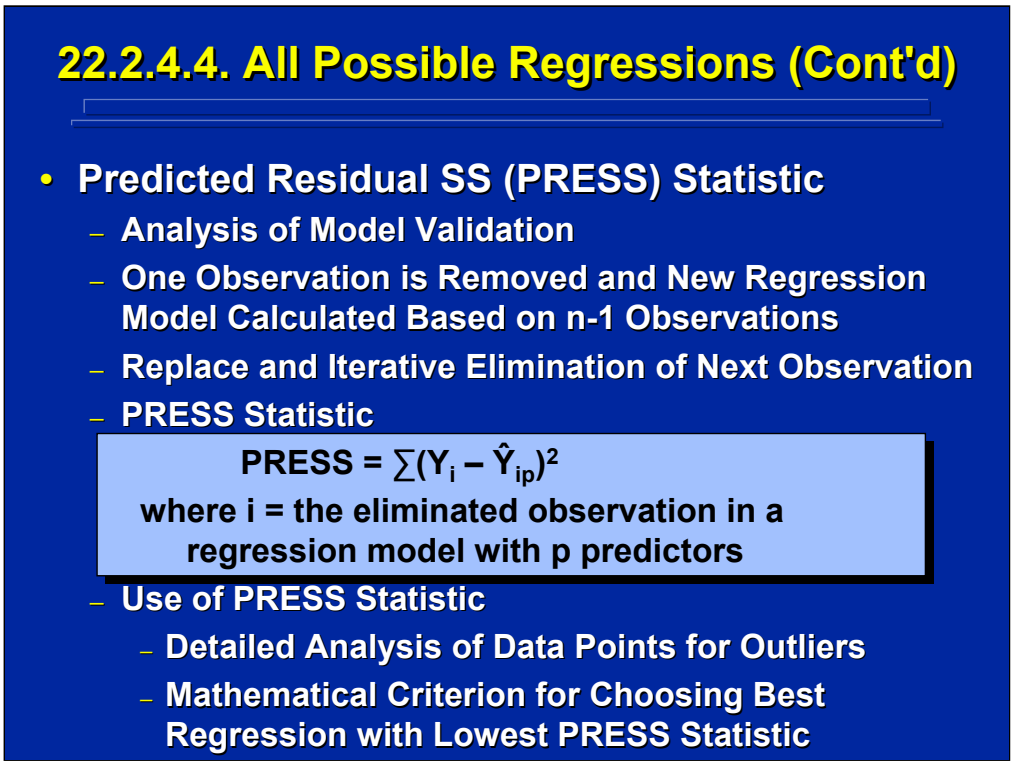

The stepwise selection procedure is a variation of forward selection in which all predictors are evaluated at each selection stage. This procedure allows for both the addition and elimination of predictors to the multiple regression equation as described in the process summarized on this slide.

Stepwise selection is often the classical regression procedure of choice because it refines forward selection and also allows backward selection. One strategy is to use all three classical regression procedures and select the consensus result as the best multiple linear regression equation.

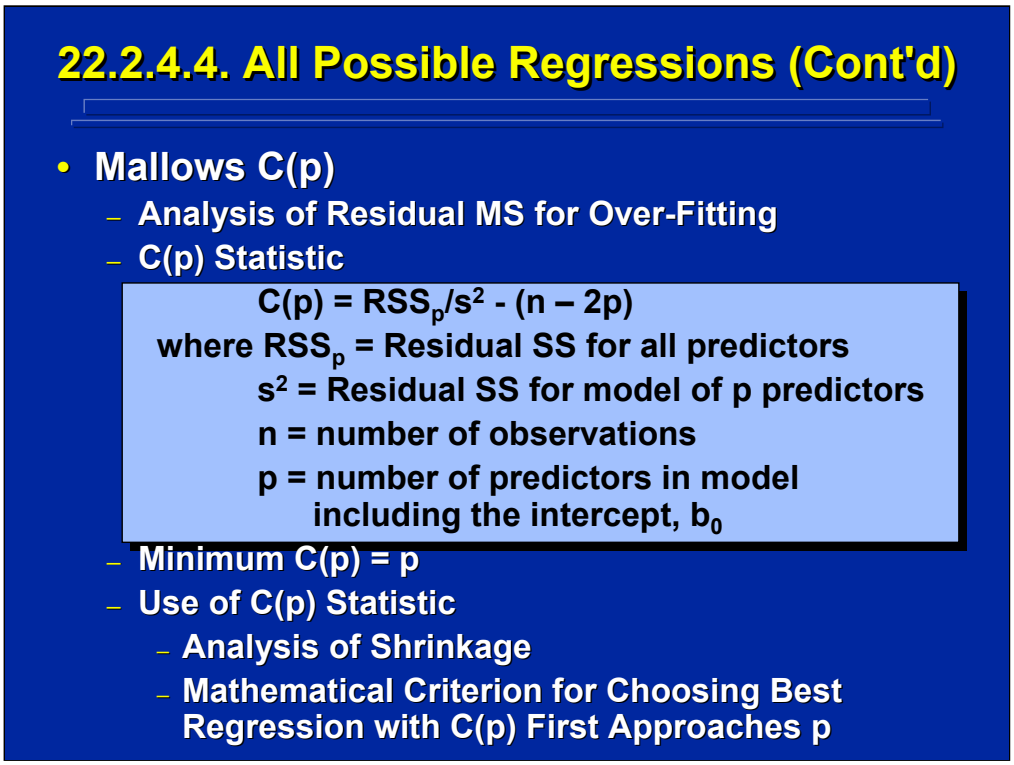



Modern regression procedures are based on conducting all possible multiple linear regression equations on the predictors and evaluating them using various statistical criteria. Usually the possible regression equations are grouped by the number of predictors and ordered by the Coefficient of Determination, $R^2$. Candidate equations with high $R^2$ in each group are then compared in terms of statistical criteria such as $R_{adj}^2$, PRESS, and Mallows C(p) goodness of fit statistics as defined in previous slides.

This selection approach can become quite cumbersome without computerized statistical analysis packages. For example, if ten predictors are considered, there are 1,023 possible multiple linear regression equations to consider as the best regression equation.




Metrics can be used to evaluate the goodness of fit of multiple linear regression when all possible regression equations are considered. One useful metric that is based directly on model validation is the PRESS statistic as described by Draper and Smith (1988, p. 325-326) and Myers (1991, pp. 170-178). The formula for the PRESS statistic is shown on the center of this slide and is based on the iterative elimination of one observation at a time in calculating a new regression equation using n-1 observations. The resulting PRESS statistic can be used to isolate data points that might be outliers in generating the regression model and can be used to choose the best regression model yielding the lowest PRESS value.




Mallows C(p) statistic is another useful modern regression procedure for estimating regression model shrinkage due to over-fitting the number of predictors (Draper and Smith, 1988, p. 298-302). The formula for Mallows C(p) provided by Draper and Smith (1988, p.299) shown on the middle portion of this slide is based on the residual sum of squares when all predictors are included in the regression model. The C(p) statistic describes the overall discrepancy (i.e. variance error and bias error) in a regression model. The value of Mallows C(p) that first approaches p, derived by changing the number of predictors, can be used to choose the best regression model.

## 22.2.5. Best Equation Example

- **Example Problem: The commander's combat operation performance in a battalion level command and control center for the Army is scored on a 100 point scale. Scores of fifteen battalion commanders are predicted as a function of four command and control tasks. The predictors are the time to complete Recognition, Decision, Communication, and, Evaluation tasks. What is the best set of significant linear predictors to use in the prediction equation ($p < 0.05$)?**

(Click in this red rectangle to see SAS calculations for this example.)

This is the same example problem based on happenstance data used previously in the multiple linear regression example. Rather than providing the multiple linear regression that includes all four predictors, this problem asks for the best possible set of the four predictors taking into account the possible covariance among predictors. See Slater and Williges (2006) for the SAS analyses related to this example problem.

## 22.2.5. Best Equation Example (Cont'd)

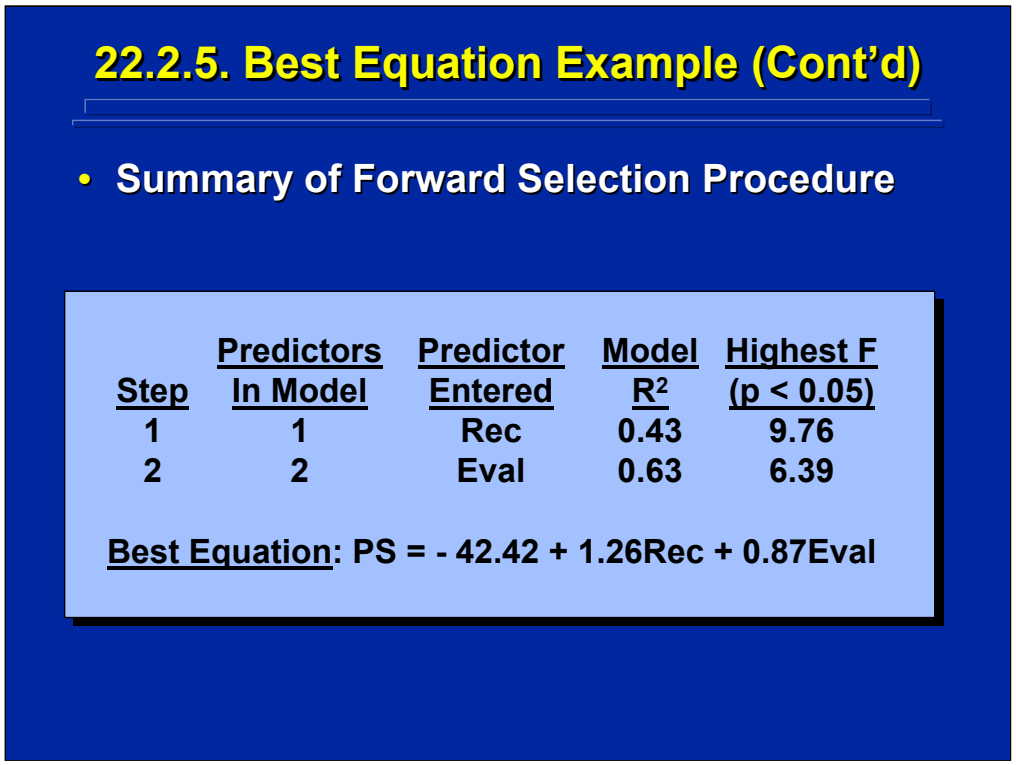
- **Summary of Backward Selection Procedure**

| Step | Predictors In Model | Predictor Deleted | Model $R^2$ | Lowest F (p>.05) |
|---|---|---|---|---|
| 1 | 3 | Com | 0.68 | 0.51 |
| 2 | 2 | Dec | 0.63 | 1.87 |

**Best Equation: PS = - 42.42 + 1.26Rec + 0.87Eval**

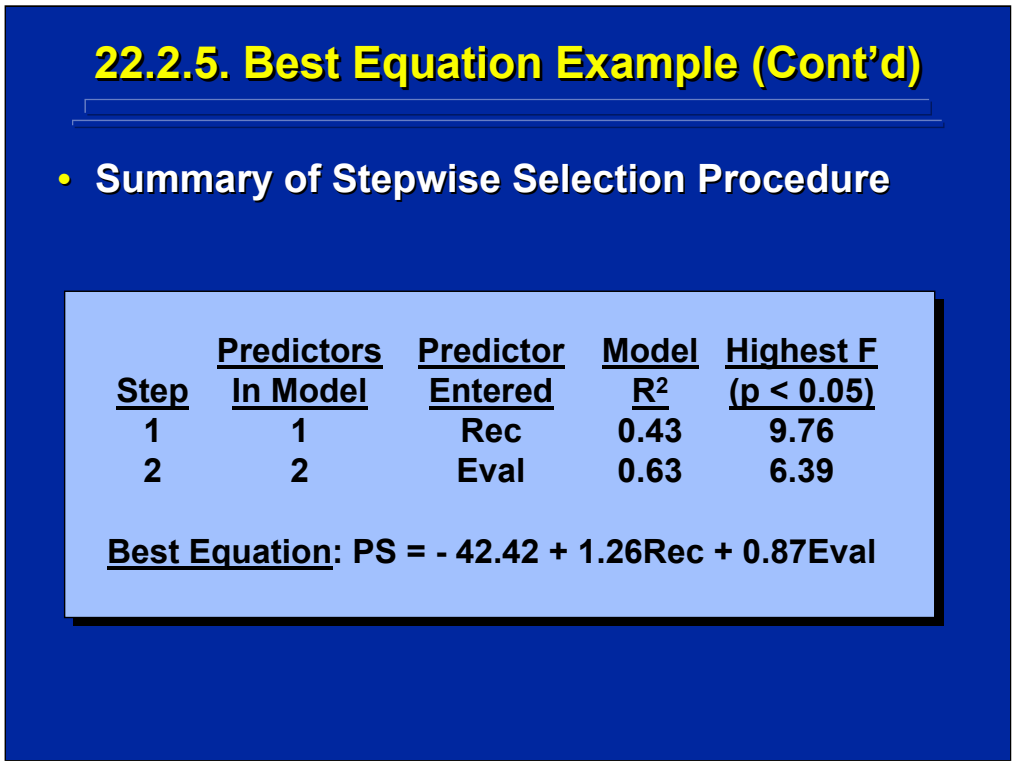
(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the results of using the backward selection procedure for choosing the best regression equation to predict a commander's overall performance score. The first predictor eliminated was the Communication subtask predictor, and the second predictor eliminated was the Decision subtask predictor. No additional predictors were deleted. Consequently, the best regression equation chosen by backward selection has two significant predictors ($p < 0.05$), the time to complete the Recognition and the Evaluation subtasks. The resulting multiple linear regression equation with least squares beta weight values is shown at the bottom of this slide.




(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the results of using the forward selection procedure for choosing the best regression equation to predict a commander's overall performance score. The first predictor added was the Recognition subtask predictor, and the second predictor added was the Evaluation subtask predictor. No additional predictors were added. Consequently, the best regression equation chosen by forward selection has two significant predictors ($p < 0.05$), the time to complete the Recognition and the Evaluation subtasks. The resulting multiple linear regression equation with least squares beta weight values is shown at the bottom of this slide.




(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the results of using the stepwise selection procedure for choosing the best regression equation to predict a commander's overall performance score. The first predictor added was the Recognition subtask predictor, and the second predictor added was the Evaluation subtask predictor. No additional predictors were added or deleted. Consequently, the best regression equation chosen by stepwise selection has two significant predictors ($p < 0.05$), the time to complete the Recognition and the Evaluation subtasks. The resulting multiple linear regression equation with least squares beta weight values is shown at the bottom of this slide.

## 22.2.5. Best Equation Example (Cont'd)

- **Summary of All Possible Regression Equations**

| Predictors in Model | R-Square | Variables in Model |
|---|---|---|
| 1 | 0.42* | Rec |
| 1 | 0.19 | Com |
| 1 | 0.13 | Eval |
| 1 | 0.01 | Dec |
| 2 | 0.62* | Rec Eval |
| 2 | 0.48 | Rec Dec |
| 2 | 0.46 | Rec Com |
| 2 | 0.27 | Com Eval |
| 2 | 0.21 | Dec Com |
| 2 | 0.15 | Dec Eval |
| 3 | 0.68* | Rec Dec Eval |
| 3 | 0.63 | Rec Com Eval |
| 3 | 0.55 | Rec DeC Com |
| 3 | 0.27 | Dec Com Eval |
| 4 | 0.70** | Rec Dec Com Eval |

*Candidate Equations within Group of Predictors

**Best Equation: PS = - 85.83 + 1.40Rec + 0.48Dec + 0.30Com + 0.78Eval

(Click in this red rectangle to see SAS calculations for this example.)

The 15 possible regression equations are grouped by one, two, three, and four predictors on this slide. The regression equations within each grouping are ordered by $R^2$. An asterisk denotes the candidate regression equations with the highest $R^2$ value in each grouping. Note that the overall highest $R^2$ is 0.70 for the regression equation with four predictors. This would be chosen as the best equation on the basis of $R^2$ and is the same as the multiple linear regression equation in the previous example. But, only the Recognition subtask predictor is a significant predictor ($p < 0.05$) in this equation. Consequently, other candidate regression equations need to be evaluated by modern regression criteria to determine the best equation.

## 22.2.5. Best Equation Example (Cont'd)

- **PRESS Statistic Evaluation of All Regression Candidates for Best Equation**

| Predictors In Model | Predictor Entered | Model $R^2$ | Adjusted $R^2$ | PRESS Statistic |
|---|---|---|---|---|
| 1 | Rec | 0.43 | 0.38 | 404.53 |
| 2 | Rec Eval | 0.63 | 0.57 | 326.64* |
| 3 | Rec Dec Eval | 0.68 | 0.59 | 413.81 |
| 4 | Re Dec Com Eval | 0.70 | 0.58 | 504.76 |

*Best Equation: PS = - 42.42 + 1.26Rec + 0.87Eval

(Click in this red rectangle to see SAS calculations for this example.)

Further evaluations of the four candidate regression equations with the highest $R^2$ within each predictor grouping are shown on this slide in terms of the $R_{adj}^2$ and the PRESS statistic. Note that the regression equations with two, three, and four predictors have essentially the same adjusted $R^2$ value (0.57. 0.58, and 0.59, respectively). But, the regression equation including the two predictors, time to complete the Recognition and Evaluation subtasks, resulted in the lowest PRESS statistic (326.64). Consequently the best regression equation based on the $R_{adj}^2$ and the PRESS statistic is the equation shown on the bottom of this slide.

## 22.2.5. Best Equation Example (Cont'd)

- **Mallows C(p) Evaluation of All Regression Candidates for Best Equation**

| Predictors In Model | Predictor Entered | Model $R^2$ | Adjusted $R^2$ | Mallows C(p) |
|---|---|---|---|---|
| 1 | Rec | 0.43 | 0.38 | 7.85 |
| 2 | Rec Eval | 0.63 | 0.57 | 3.30* |
| 3 | Rec Dec Eval | 0.68 | 0.59 | 3.51 |
| 4 | Re Dec Com Eval | 0.70 | 0.58 | 5.00 |

*Best Equation: PS = - 42.42 + 1.26Rec + 0.87Eval

(Click in this red rectangle to see SAS calculations for this example.)

Further evaluations of the four candidate regression equations with the highest $R^2$ within each predictor grouping are shown on this slide in terms of the $R_{adj}^2$ and the Mallows C(p) criteria. As noted on the previous slide, the regression equations with two, three, and four predictors have essentially the same adjusted $R^2$ value (0.57. 0.58, and 0.59, respectively). But, the regression equation including the two predictors, time to complete the Recognition and Evaluation subtasks, resulted in the Mallows C(p) value (3.30) first approaching p. Consequently, the best regression equation based on the $R_{adj}^2$ and the Mallows C(p) statistic is the equation shown on the bottom of this slide.

## 22.2.5. Best Equation Example (Cont'd)

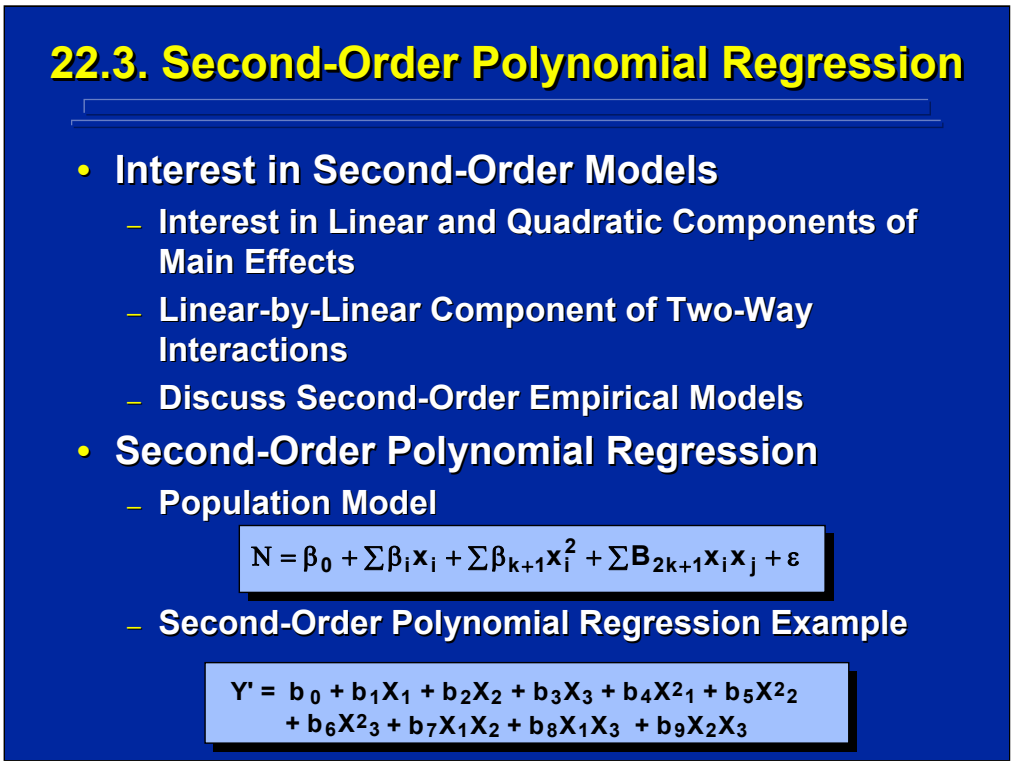
- **Tradeoff of All Procedures and Criteria**
  - Same Result for All Classical Selection Procedures
  - High $R^2$ of Candidate Equations in All Regressions
  - Low Shrinkage in $R^2_{Adj}$ of Candidate Regressions
  - Lowest PRESS Statistic of Candidate Regressions
  - Lowest Mallows C(p) of Candidate Regressions
- **Best Regression Equation Selected**

**Best Equation: PS = - 42.42 + 1.26Rec + 0.87Eval**

| Model $R^2$ | Adjusted $R^2$ | PRESS Statistic | Mallows C(p) |
|---|---|---|---|
| 0.63 | 0.57 | 326.64 | 3.30 |

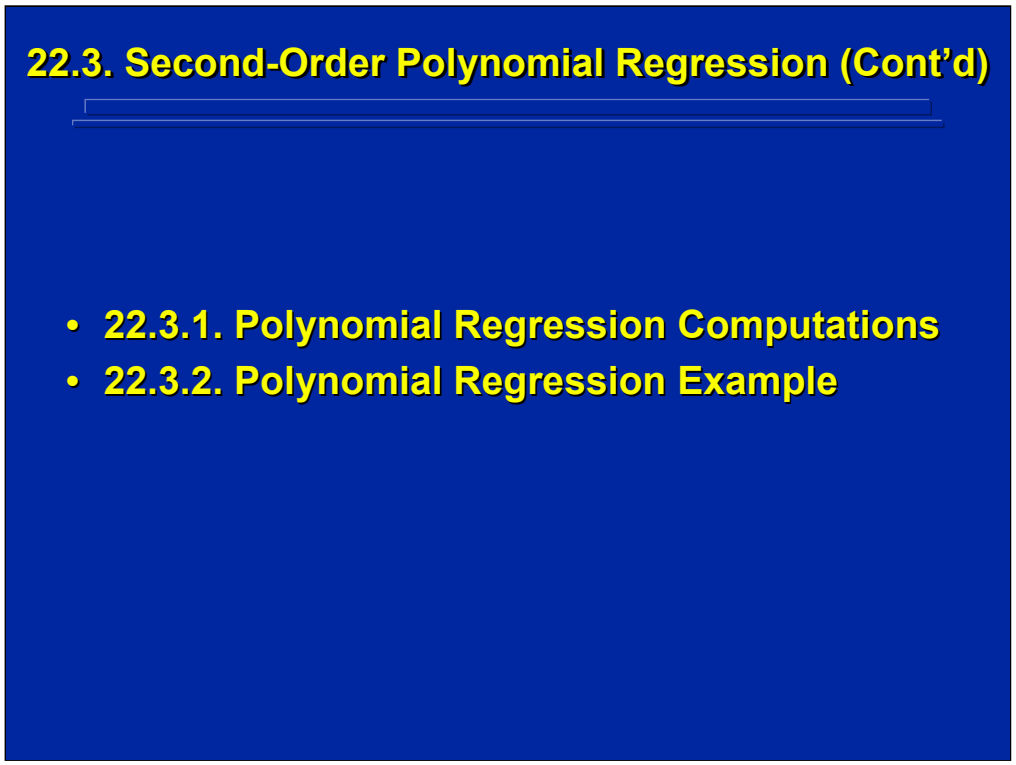
(Click in this red rectangle to see SAS calculations for this example.)

In summary, there are a variety of procedures and criteria that can be considered in choosing the best set of predictors to include in the multiple linear regression equation. Based on a tradeoff of these techniques, the best regression equation for this example problem is the two predictor equation shown at the bottom of this slide. All the classical selection procedures result in this equation. Based on an evaluation of candidate equations resulting from all possible regression equations including one to four predictors this two predictor regression equation has the lowest PRESS statistic, the Mallows C(p) value that first approaches p, a high Coefficient of Determination (0.63), and a low estimated shrinkage in the Coefficient of Determination (0.57). Consequently, the best regression equation for predicting a commander's overall performance score is the one with the two significant predictors ($p < 0.05$), the time to complete the Recognition and the time to complete the Evaluation subtasks.

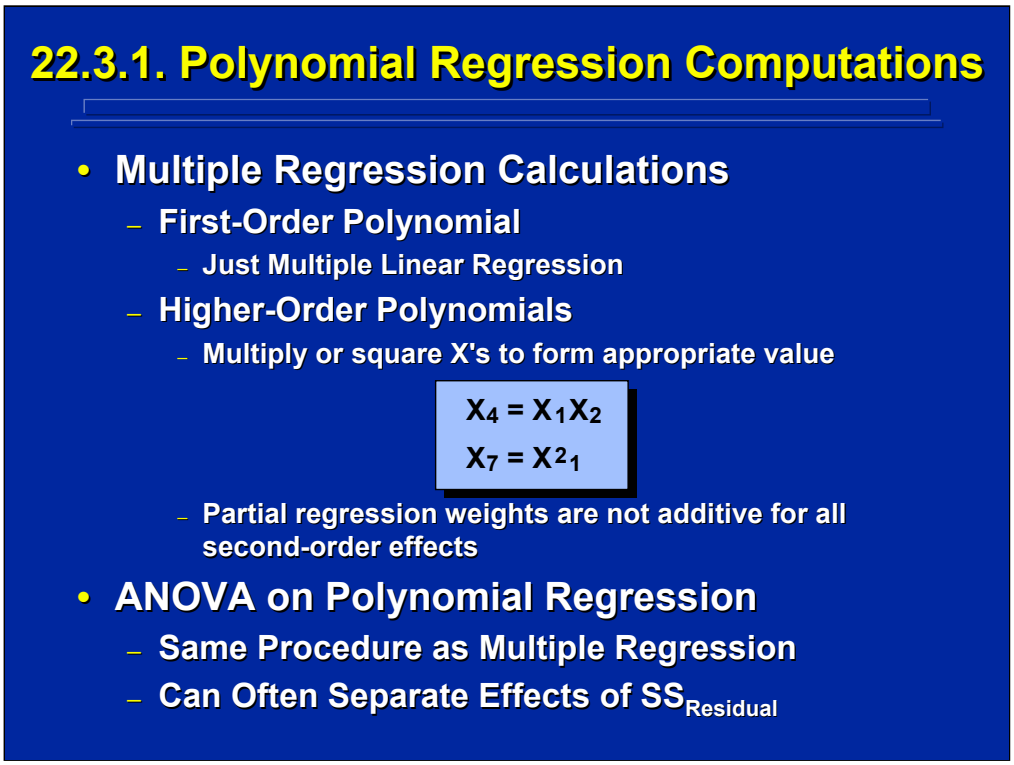



In human factors research, a polynomial expression is a convenient way to represent a variety of underlying relationships and can form the basis of empirical models to predict human performance in complex systems. Usually only first-order and second-order empirical models are used because they cover most human behavior effects. Consequently, the human factors researcher should plan to collect enough data to generate up to a complete second-order polynomial empirical model plus some extra data to test model lack of fit due to the existence of higher-order effects.

The general form of a complete second-order polynomial regression model is shown on the bottom portion of this slide. The population model shows the linear, pure quadratic, and linear-by-linear effects of predictors. An example of a complete second-order polynomial regression which includes three factors is shown at the bottom of this slide.




This subsection describes the general procedures for computing a second-order polynomial regression and provides an example of conducting polynomial regression using data from a 2x3 factorial ANOVA design.




Two general analyses are conducted in polynomial regression using computerized statistical packages. First, the multiple regression analysis is conducted to determine the line of best fit. First-order polynomial regression analysis is the same as multiple linear regression analysis covered in the previous section. Higher-order polynomial regression analysis uses multiple linear regression analysis, but the Xs forming the higher-order terms are multiplied together or squared first in order to generate the X term used in the multiple linear regression as shown in the middle portion of this slide.

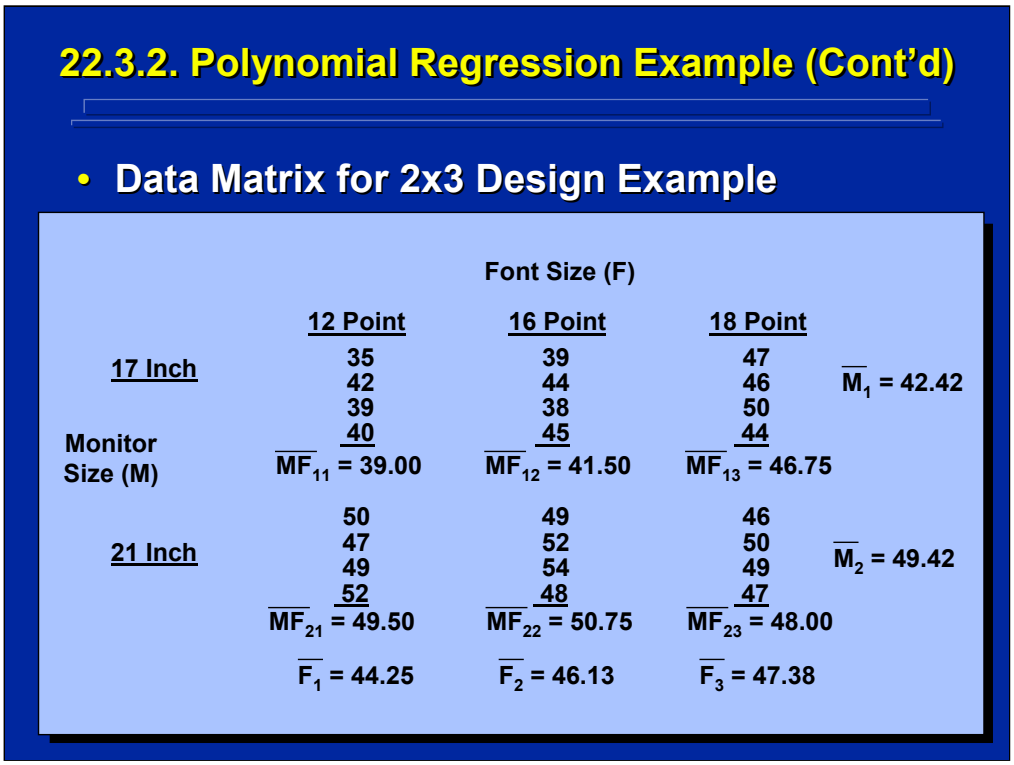
Second, an ANOVA is conducted on the polynomial regression to determine the goodness of fit. This procedure is the same as used in multiple linear regression demonstrated in the previous section. Often it is possible to separate the $SS_{Residual}$ into error and lack of fit for testing higher-order effects not included in the polynomial regression model. In SAS, for example, the response surface regression procedure can be used to test the goodness of fit of first-order and second-order effects and lack of fit in second-order polynomial regression models.

## 22.3.2. Polynomial Regression Example

- **Example Problem: A between-subjects experiment (n = 4) was conducted to build an empirical model of soldier percent reading comprehension of text presented on computer displays as a function of possible first- and second-order effects involving two different sizes of computer monitors (17 and 21 inch) and three different font sizes (12, 16, and 18 point). What is the resulting second-order model and were any first- and second-order parameters significant predictors (p < 0.01)?**

(Click in this red rectangle to see SAS calculations for this example.)

This example problem is a between-subjects 2x3 factorial design that provides the data to generate the second-order polynomial regression model to predict percent reading comprehension as a function of two predictors, computer monitor size and font size. The SAS programs and descriptions of the computer analyses of this example problem are provided in the Slater and Williges (2006) appendix.




(Click in this red rectangle to see SAS calculations for this example.)

This page provides the hypothetical data from the 2x3 factorial design described on the previous slide based on a sample size of four observations. The means for the two main effects and the two-way interaction are also listed on the slide.

## 22.3.2. Polynomial Regression Example (Cont'd)

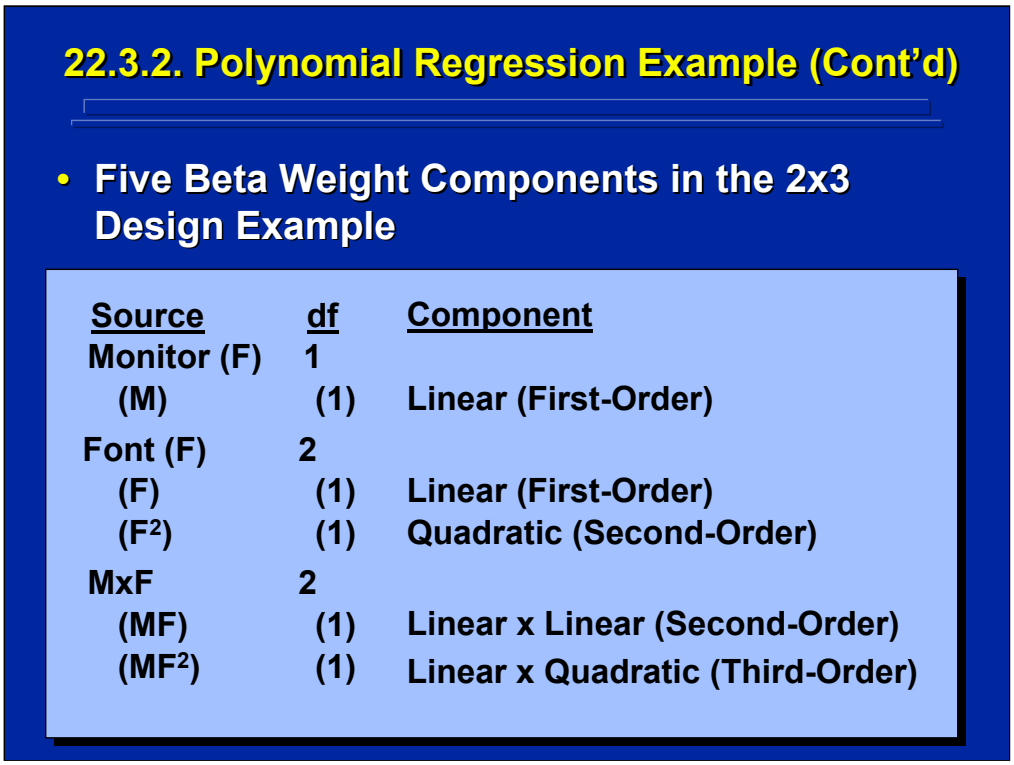
- **ANOVA Summary Table for 2x3 Design Example**

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Monitor (M) | 1 | 294.00 | 294.00 | 41.53 ** |
| Font (F) | 2 | 39.58 | 19.79 | 2.80 |
| MxF | 2 | 100.75 | 50.38 | 7.12 * |
| Subjects/MF | 18 | 127.50 | 7.08 | |
| Total | 23 | 561.83 | | |

*p < 0.01 **p < 0.0001

(Click in this red rectangle to see SAS calculations for this example.)

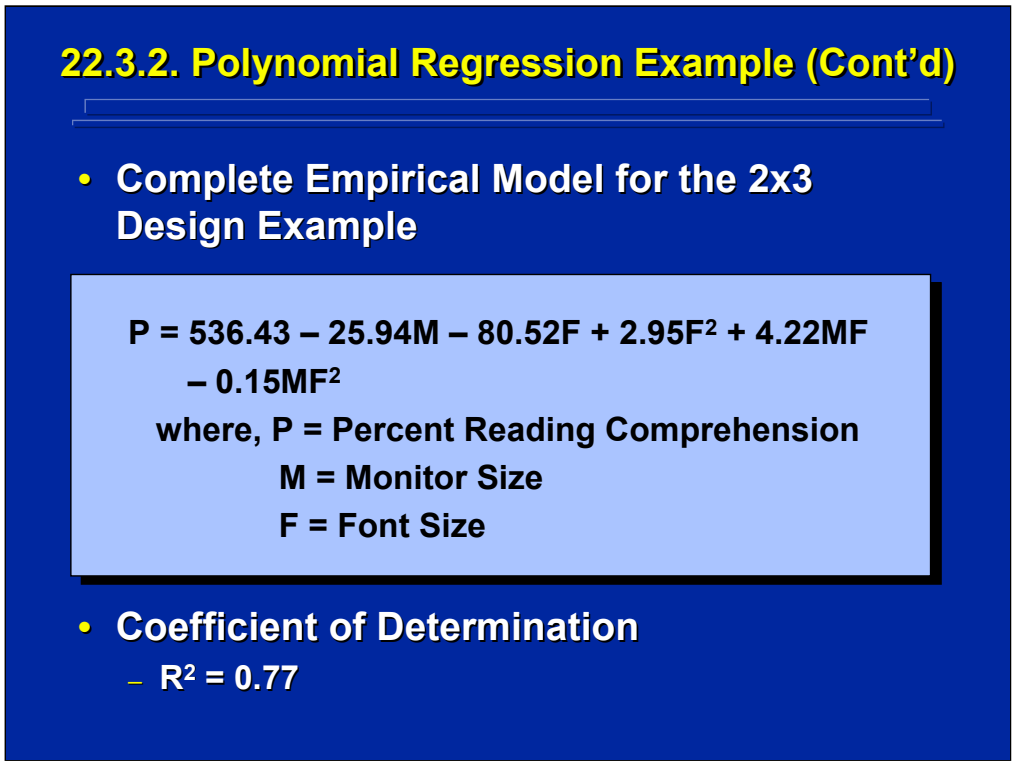
This slide provides the ANOVA Summary Table of the between-subjects 2x3 factorial design showing that both the main effect of computer monitor size and the monitor size by font size interaction are significant at the 0.01 level at least. The means for these effects are shown on the previous slide. As expected, mean percent reading comprehension was greater using the 21" computer monitor (49.42) than when using the 17" computer monitor (42.42). Additional post hoc analyses are required to isolate the significant two-way interaction.




(Click in this red rectangle to see SAS calculations for this example.)

This example problem asks for the second-order empirical model that predicts percent reading comprehension as a function of computer display size and font size rather than the significant effects in the 2x3 factorial design. In order to develop a polynomial regression empirical model, the experimenter must first determine the various one degree of freedom beta weights that are present in the 2x3 factorial design data set.

The breakdown of the five possible beta weights is shown on this slide for the example 2x3 factorial design. Note that there are two first-order effects (M and F), two second-order effects ($F^2$ and MF) and one third-order effect ($MF^2$) in this 2x3 factorial design data set that can be used as predictors in the polynomial regression empirical model.




(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the polynomial regression equation that includes all five 1 df beta weights that can be fit using the 2x3 factorial design data. Seventy-seven percent of the variation in the data can be accounted for by this empirical model.

## 22.3.2. Polynomial Regression Example (Cont'd)

- **Regression ANOVA on the Complete Empirical Model for the 2x3 Design Example**

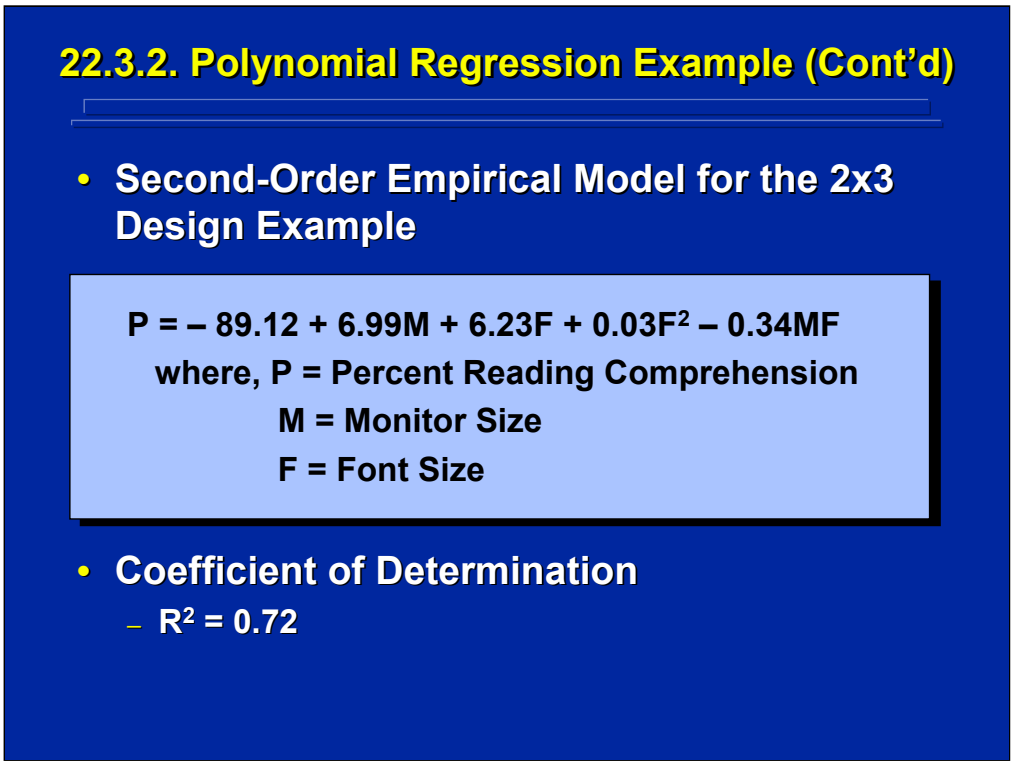
| Source | df | SS | MS | F |
|---|---|---|---|---|
| Regression Model | (5) | (434.33) | (86.87) | (12.27)** |
| M | 1 | 294.00 | 294.00 | 41.53 ** |
| F | 1 | 39.36 | 39.36 | 5.56 |
| $F^2$ | 1 | 0.22 | 0.22 | 0.03 |
| MF | 1 | 69.67 | 69.67 | 9.84 * |
| $MF^2$ | 1 | 31.08 | 31.08 | 4.39 |
| Regression Error (S/MF) | 18 | 127.50 | 7.08 | |
| Total | 23 | 561.83 | | |

*p < 0.01 **p < 0.0001

(Click in this red rectangle to see SAS calculations for this example.)

This slide presents the ANOVA Summary Table for testing the significance of the overall regression model and each of the partial regression weights for the empirical model shown on the previous slide. The partial F-test on each beta weight assumes all the other predictors in the regression model are present. Consequently, the total of the sum of squares for all the partial regression weights equals the sum of squares of the regression model (434.33).

Regression error is used as the error term for each F-test. Since this is a complete model, the sum of squares due to regression error is the same as the sum of squares for S/MF used in the ANOVA of the 2x3 factorial design summarized in a previous slide. Note that the overall model, and the linear effect of monitor size (M), and the linear-by-linear component of the display size-by-font size interaction (MF) are significant predictors of percent reading comprehension at the 0.01 level of significance.




(Click in this red rectangle to see SAS calculations for this example.)

The example problem asked for the second-order empirical model, not the complete model that could be determined by the 2x3 factorial design data. The requested second-order polynomial regression equation is shown on this slide. Note that it does not include the third-order partial regression weight due to the linear-by-quadratic component ($MF^2$) of the two-way interaction of monitor size and font size. Only 72% of the variation is accounted for by this empirical model as compared to 77% of the variation accounted for by the complete polynomial equation shown on a previous slide.

## 22.3.2. Polynomial Regression Example (Cont'd)

- **Regression ANOVA on the Second-Order Empirical Model for the 2x3 Design Example**

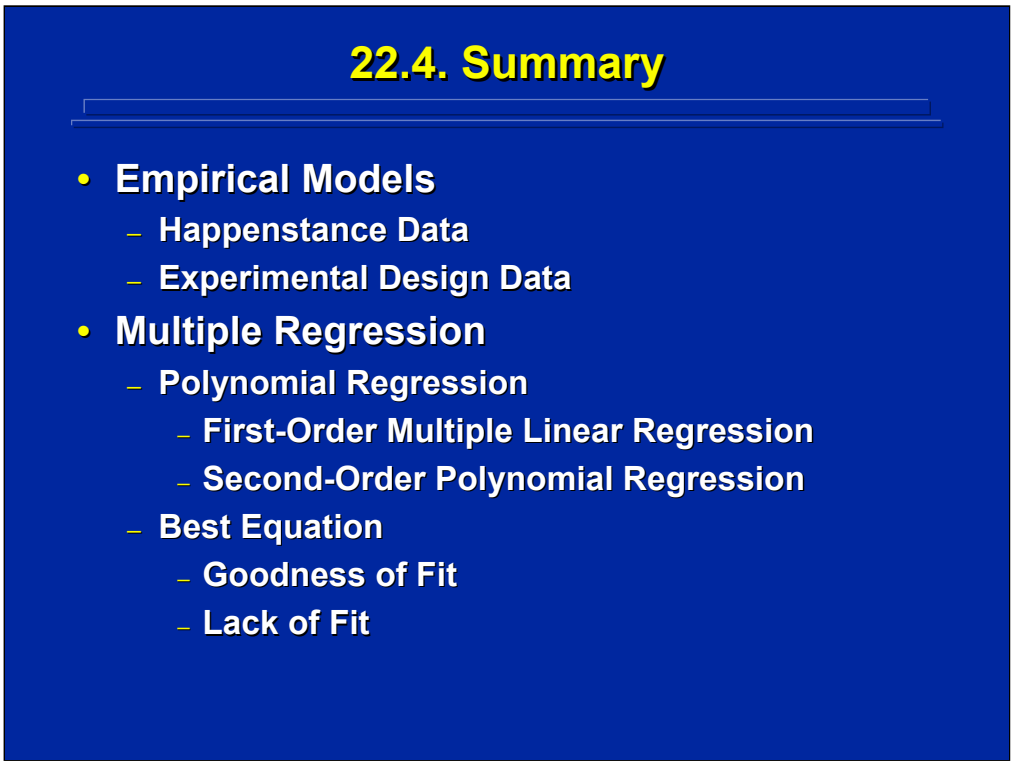
| Source | df | SS | MS | F |
|---|---|---|---|---|
| Regression Model | (4) | (403.25) | (100.81) | (12.08)** |
| M | 1 | 294.00 | 294.00 | 35.21 ** |
| F | 1 | 39.36 | 39.36 | 4.71 |
| $F^2$ | 1 | 0.22 | 0.22 | 0.03 |
| MF | 1 | 69.67 | 69.67 | 8.34 * |
| Regression Error | (19) | (158.58) | (8.35) | |
| Lack of Fit ($MF^2$) | 1 | 31.08 | 31.08 | 4.39 |
| Refined Error (S/MF) | 18 | 127.50 | 7.08 | |
| Total | 23 | 561.83 | | |

*p < 0.01 **p < 0.0001

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the ANOVA Summary Table for testing the significance of the overall regression model and each of the partial regression weights for the empirical model shown on the previous slide. The partial F-test on each beta weight assumes all the other predictors in the regression model are present. Therefore, the total of the sum of squares for all the four partial regression weights equals the sum of squares of the regression model (403.25). The fifth partial regression weight, $MF^2$, is not included in the regression, but is listed as Lack of Fit (LOF) under error. Consequently, regression error is pooled and equals LOF plus the error used in the complete model (S/MF).

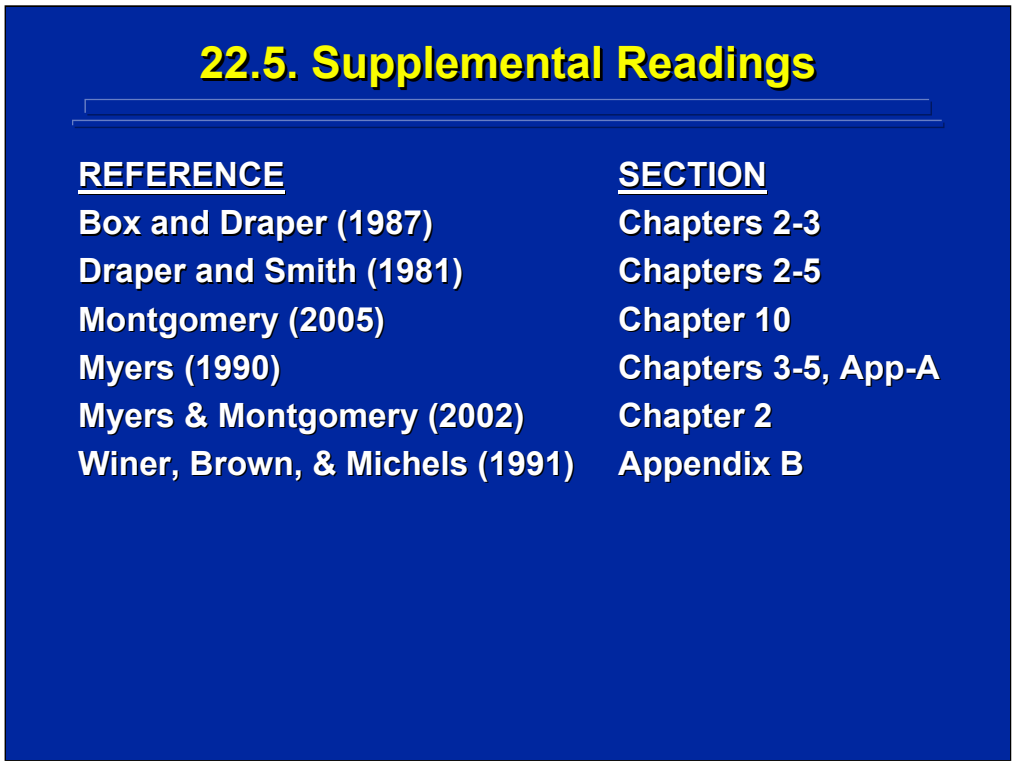
The pooled regression error is used as the error term for each F-test in the empirical model since LOF was not significant (p>0.01) when tested by Refined Error. Alternatively, the Refined Error could be used as the error term for all F-tests and would provide the F-ratios used in the complete model. Note that the overall model, the linear effect of monitor size (M), and the linear-by-linear component (MF) of the display size-by-font size interaction are significant predictors of percent reading comprehension at the 0.01 level of significance.




By way of summary, empirical models including multiple predictors can be generated through multiple regression by using either happenstance or experimental design data. Experiments provide more control and are more efficient in collecting data for empirical models in human factors.

Polynomial regression is the general form of multiple regression that can include both linear effects and higher-order effects. Polynomial regression analysis includes procedures for determining both the line of best fit and the goodness of fit of the regression equation. Multiple linear regression is the same as a first-order polynomial regression and uses the least squares criterion for determining the line of best fit. For most human factors applications, first- and second-order polynomials account for most aspects of human performance in complex systems.

Both classical and modern regression procedures can be used to determine the best multiple regression equation when the predictors are correlated as often occurs with happenstance data. Various statistics such as $R^2$, $R_{Adj}^2$, PRESS, and Mallows C(p), as well as tests of significance of both the regression model and the partial regression weights, are used to evaluate the goodness of fit of the multiple regression. Evaluation of regression lack of fit can be used to determine the possible need for higher-order effects in the empirical model.



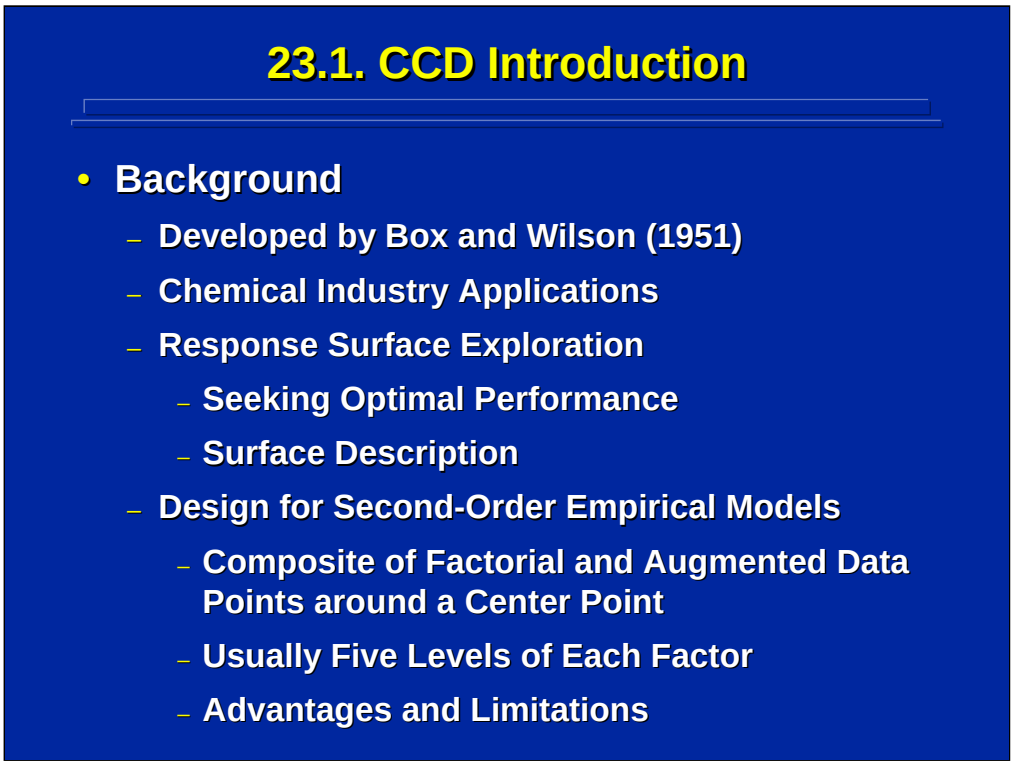

The first four texts listed on this slide provide a general discussion of multiple regression. The Draper and Smith (1981), Myers (1990), and Winer, et al. (1991) texts review matrix algebra as used in regression analysis. Box and Draper (1987) and Myers and Montgomery (2002) describe polynomial regression applications to empirical model building.

# Topic 23. Central-Composite Designs (CCD)

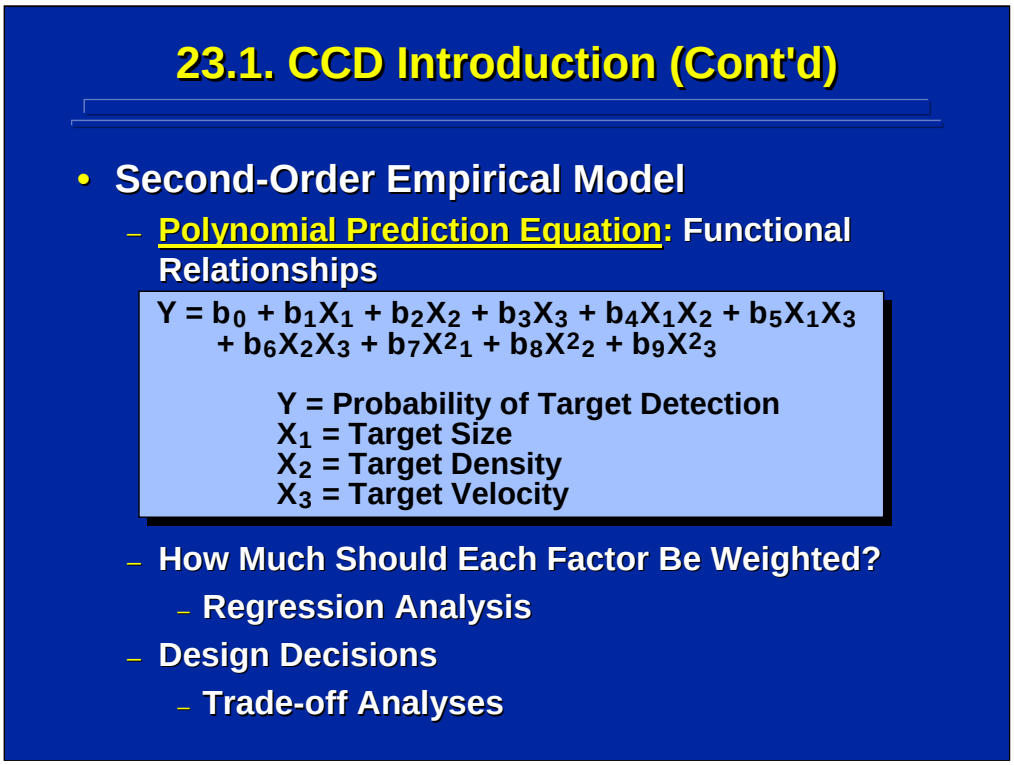
23.1. CCD Introduction
23.2. CCD Specification
- 23.2.1. CCD Configuration
- 23.2.2. Replication
- 23.2.3. Value of $\alpha$

23.3. CCD Analysis
23.4. CCD Examples
23.5. Alternative Second-Order Designs
23.6. Summary
23.7. Supplemental Readings

Topic 23 describes experimental designs that can be used to collect data for building second-order empirical models. Specifically, this topic focuses on central-composite designs (CCD) that were developed to explore response surfaces using empirical models. The background, specification, analysis, and an example of a CCD along with a comparison to alternative second-order experimental designs are discussed in this topic. Finally, a summary along with supplemental readings on CCD in current experimental design textbooks is provided.



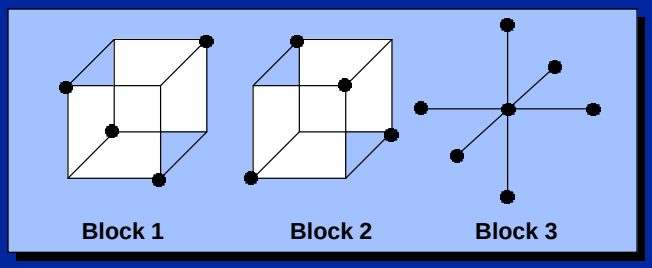

The CCD was developed by Box and Wilson (1951) as part of response surface methodology for seeking optimum yield of chemical compounds. The CCD was specifically developed as an efficient data collection procedure for fitting second-order empirical models through sequential experiments. The design is a composite of $2^k$ or $2^{k-p}$ factorial points and augmented data points around a center point that usually yields five different levels of each factor in the experimental design. Hence, the design is named a central-composite design. The CCD has several mathematical advantages for developing and testing the adequacy of empirical models and some disadvantages for building global prediction equations. Williges and Simon (1971) provide a detailed discussion of several advantages and limitations of CCD for human factors and ergonomics research.




Second-order models developed from CCD data are polynomial regression prediction equations that predict performance as a function of several quantitative predictors (i.e., factors). The target detection example shown on this slide is a second-order polynomial with three predictors of target size, density, and velocity. The empirical values of each partial regression weight, $b_i$, are least squares criterion solutions to the polynomial regression using data from the CCD. The resulting prediction equation can be used to predict target detection, determine the relative weights of the predictors, and assist in design tradeoffs of parameters in complex systems instead of just testing the statistical significance of various effects defined by the three target detection factors. Not only does the CCD have the advantage of providing the necessary and sufficient data to fit a second-order empirical model, the CCD also provides additional data to test the adequacy of the fit of the empirical model.

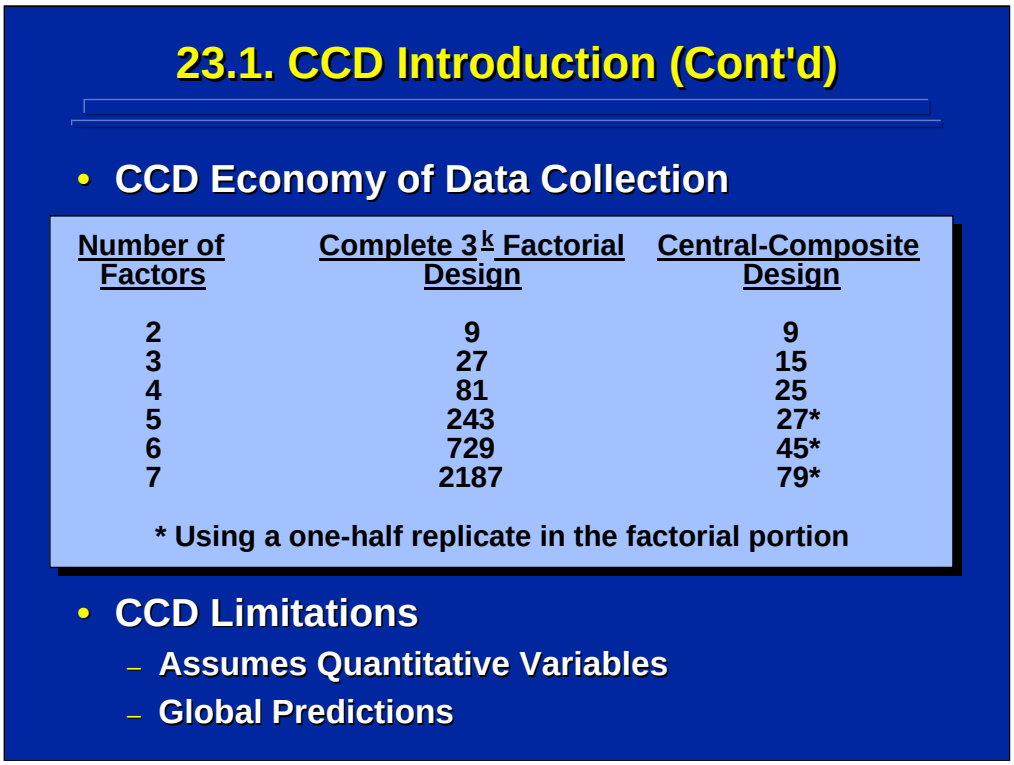
## 23.1. CCD Introduction (Cont'd)

- CCD Blocking



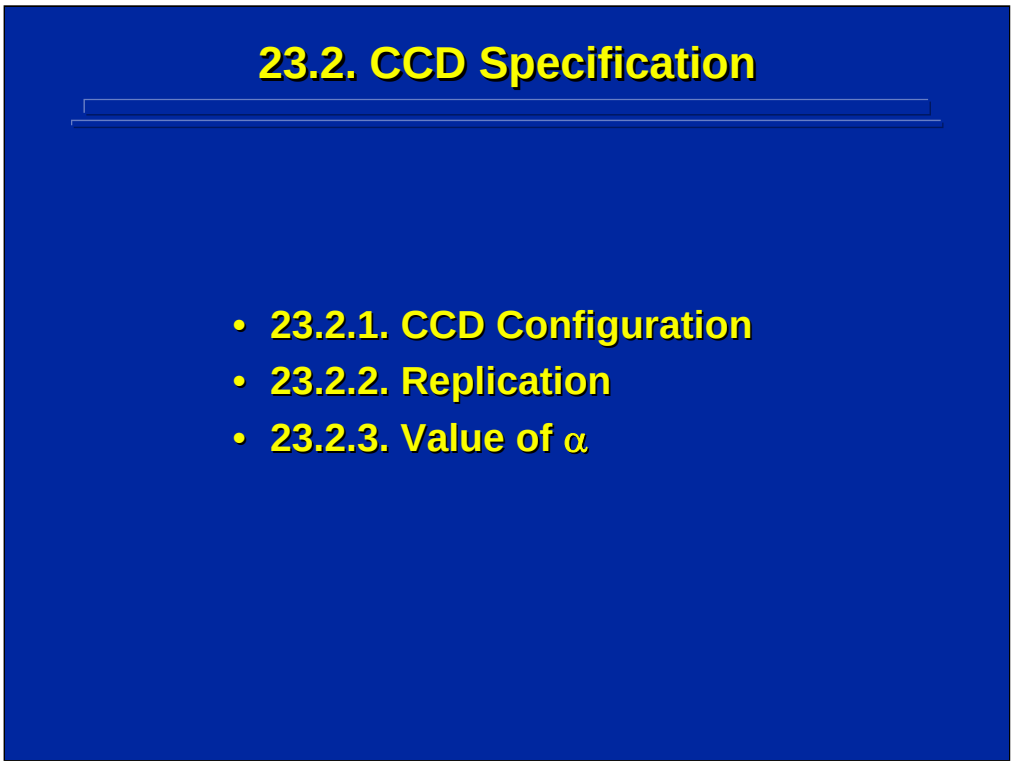

  - Control of Nuisance Variables
  - Flexibility to Collect Data in Stages
    - Add or Drop Variables
    - Test Order of Polynomial

Another important advantage of the CCD is that the data can be collected in blocks if the experimenter chooses to do so. For example, the three-factor CCD data points on this slide are divided into three orthogonal blocks. Blocking allows the experimenter to control nuisance variables, such as data collection days, by keeping any effect of the nuisance variable orthogonal to the empirical model. Blocking the CCD design also allows for data collection in stages for making decisions to add and drop factors included in the empirical model or for determining if more data are needed for a higher-order empirical model.

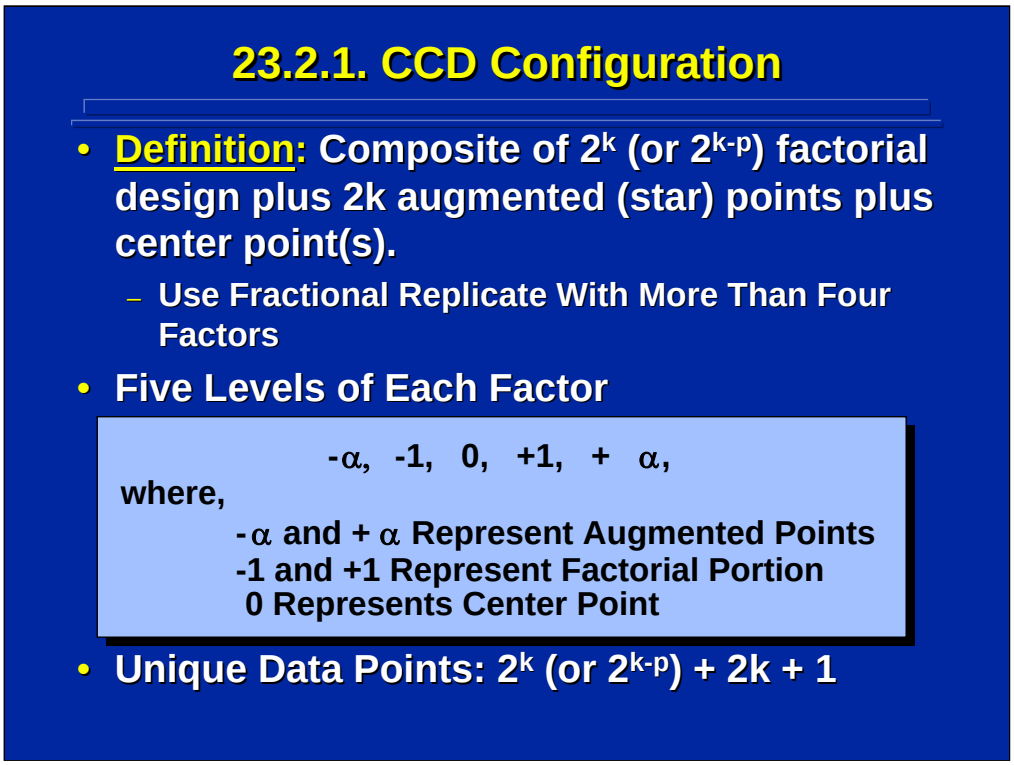



A major advantage of the CCD is its economy in data collection for fitting and testing the adequacy of second-order empirical models. A minimum of three levels of each factor must be observed to fit a second-order empirical model. The top portion of this slide lists the unique data points of a CCD as compared to its $3^k$ factorial design counterpart. When more than two factors are included in the empirical model, the CCD is more economical than the $3^k$ factorial design because the factorial design provides data to test higher-order effects rather than just first- and second-order effects.

As shown on the bottom of this slide a CCD is not without limitations. First, all the factors included in the CCD are assumed to be quantitative in order to set the precise levels defined by the CCD configuration. Second, if the empirical model is used for global prediction across the entire effective range of each factor, the spacing between levels in the CCD may not provide adequate coverage for determining reliable global empirical models.



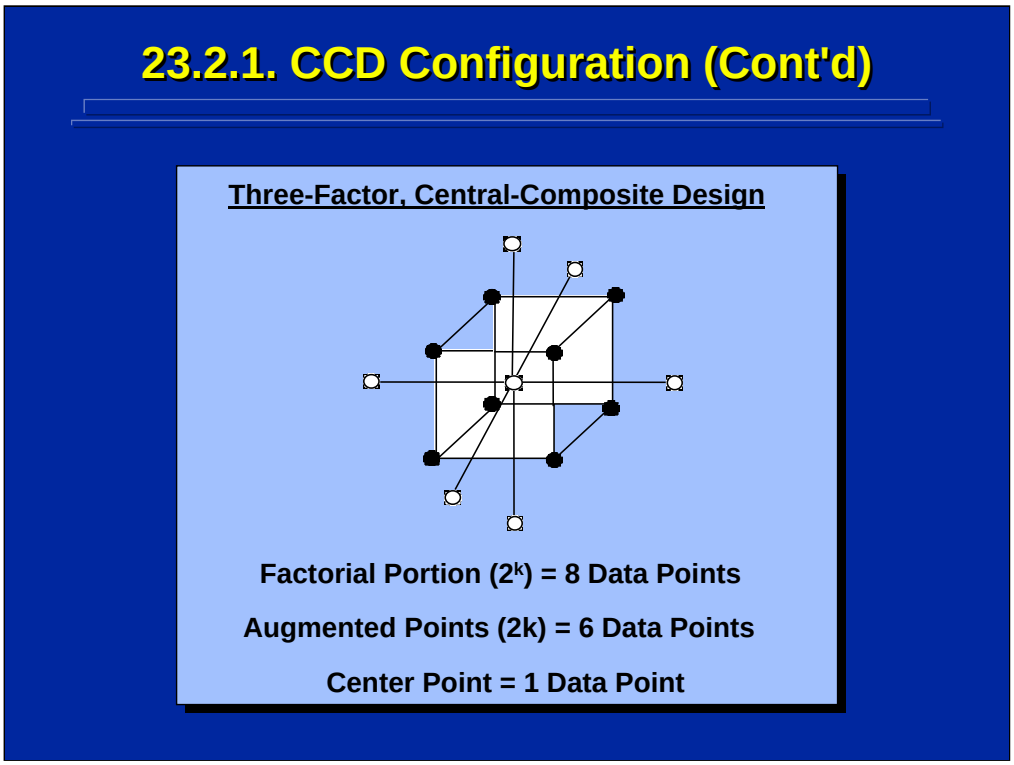

This sub-section is devoted to describing the complete specification of a CCD. The coded design configuration, choices in design replication, and calculation of the $\alpha$ coded value are described separately.




The definition of a CCD is provided at the top of this slide. Basically a CCD is a composite of a $2^k$ factorial design and star points radiating from a center point. A Resolution V fractional replicate of the $2^k$ design portion of the CCD is used when five or more factors are considered in the empirical model.

The CCD is specified in general terms by using five coded values as shown in the center of this slide. The $\pm \alpha$ represents the star portion, the $\pm 1$ values represent the factorial portion, and 0 represents the center point of the CCD. Linear transformations are made between these coded values and real-world values of actual factors used in the CCD experiment.

The factorial portion of the CCD has $2^k$ or $2^{k-p}$ data points, the star portion of the CCD has each level of $\pm\alpha$ appearing at the 0 level of the other factors yielding 2k data points, and the center point of the CCD is defined by the 0 level of all factors. In general, the total number of unique data points in any CCD is shown on the bottom of this slide. For example, a three-factor CCD would have 15 unique data points (i.e. $2^3 + 2*3 + 1$).




This slide depicts a geometric representation of the 15 unique data points in the example three-factor CCD described on the previous slide. Note that 14 of the data points radiate around the center point shown as a white circle. The factorial portion of the CCD forms a cube, and the data points of this portion of the design are shown as black circles. The data points of the star portion of the CCD are shown as white circles radiating from the center of each of the six surfaces of the cube.

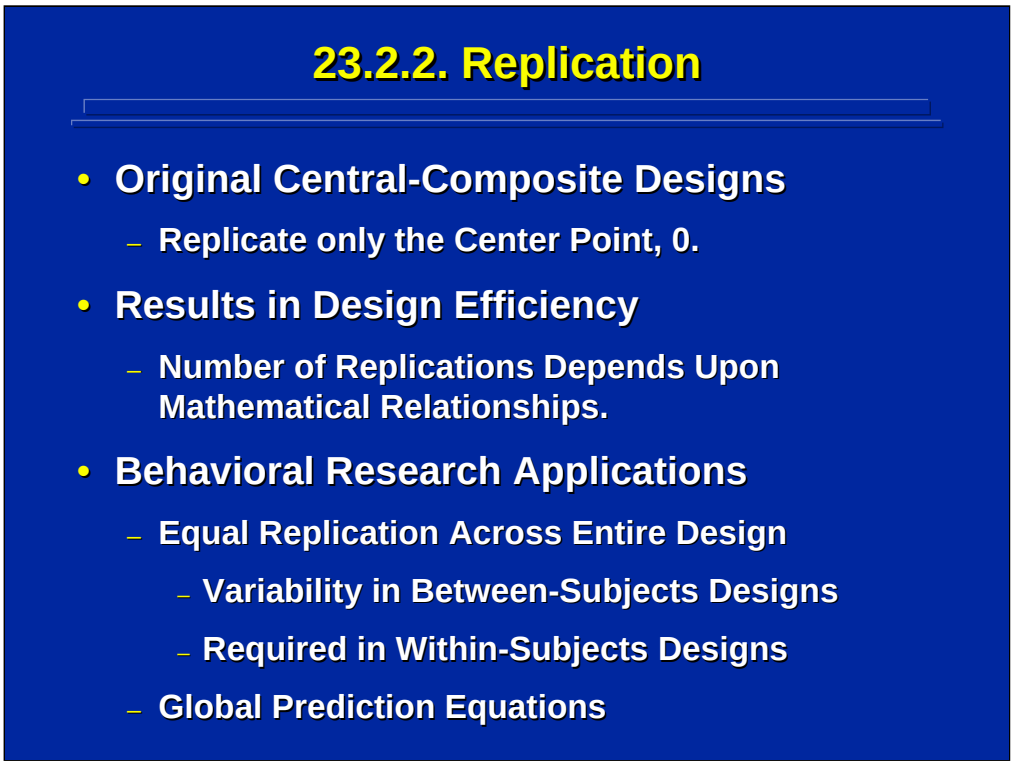
## 23.2.1. CCD Configuration (Cont'd)

**Three-Factor, Central-Composite Design**

| Treatments | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| 1 | +1 | +1 | +1 |
| 2 | +1 | +1 | -1 |
| 3 | +1 | -1 | +1 |
| 4 | +1 | -1 | -1 |
| 5 | -1 | +1 | +1 |
| 6 | -1 | +1 | -1 |
| 7 | -1 | -1 | +1 |
| 8 | -1 | -1 | -1 |
| 9 | $+\alpha$ | 0 | 0 |
| 10 | $-\alpha$ | 0 | 0 |
| 11 | 0 | $+\alpha$ | 0 |
| 12 | 0 | $-\alpha$ | 0 |
| 13 | 0 | 0 | $+\alpha$ |
| 14 | 0 | 0 | $-\alpha$ |
| 15 | 0 | 0 | 0 |

This slide specifies the 15 data points of the three-factor CCD shown geometrically in the previous slide in terms of the coded values of each of the three factors for each of the 15 data points. The first eight data points are defined by the $2^3$ factorial design in terms of combinations of the ± 1 coded levels of each factor. The next six data points show the treatment combinations of the star portion of the CCD. Note that these data points are the ± $\alpha$ coded levels of one factor in combination with the 0 coded levels of the other two factors. The last data point is the center point shown as the 0 coded level of each of the three factors.

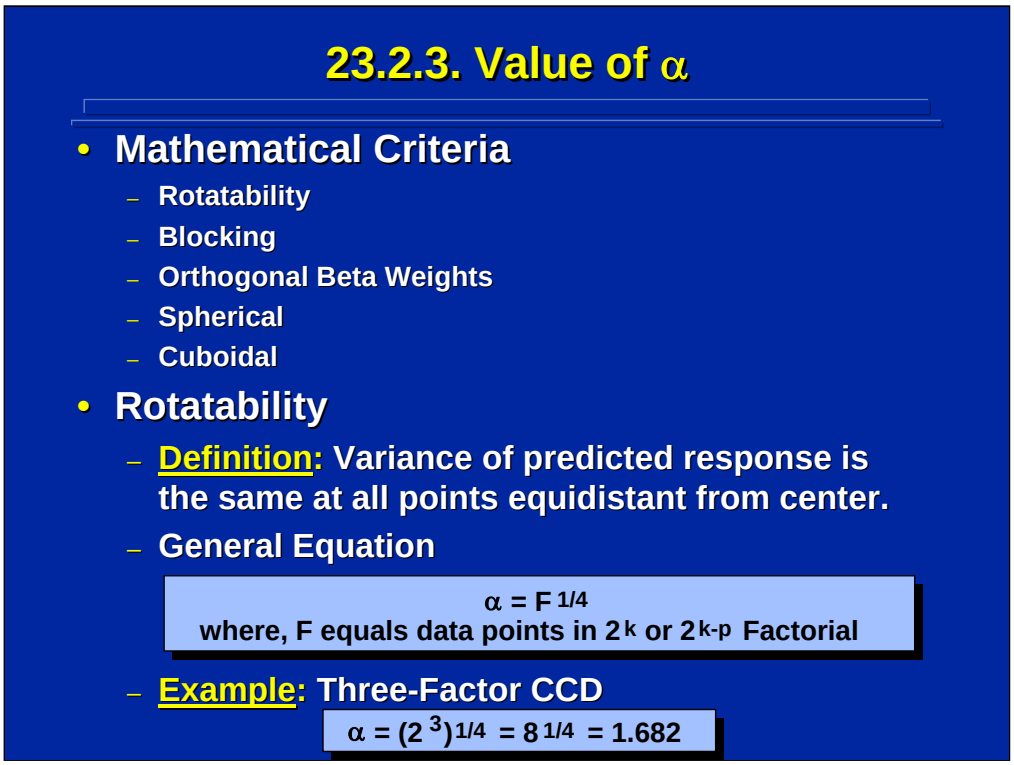
Three steps are required to translate these coded values into treatment combinations specified by the real world factor levels. First, determine the number of data points as shown on this slide. Second, specify the coded value of $\alpha$. Third, translate the coded design into the real-world values of the factors being investigated using a linear transformation.




Replication to estimate error variance in CCD usually occurs only at the center point of the design to minimize data collection. The exact number of center point replications required can depend upon the various characteristics of the design as described by Myers and Montgomery (2002).

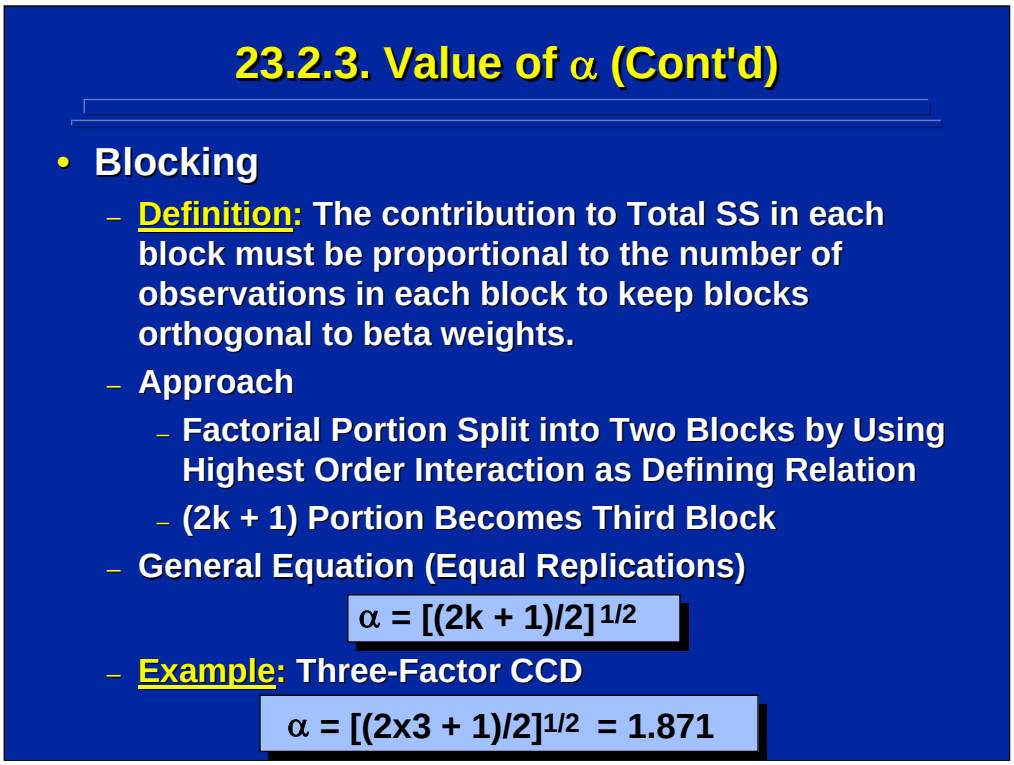
Clark and Williges (1973) recommended using equal replication across each data point in the CCD as done in experimental designs described in previous topics in this reference material. If the CCD is a between-subjects design, variability may differ across data points and a pooled estimate of error variance may be more accurate than just error estimated at the center of the design. If the CCD is a within-subjects design, every subject must receive every treatment condition that results in equal replication across the design by definition.

Even though equal replication is not as economical in terms of data collection as the original CCD that has only center point replication, the additional data collection may be warranted when generating an empirical model for global prediction of human performance where the range across levels may be large. Consequently, only CCDs with equal replication across data points are described in this reference material.



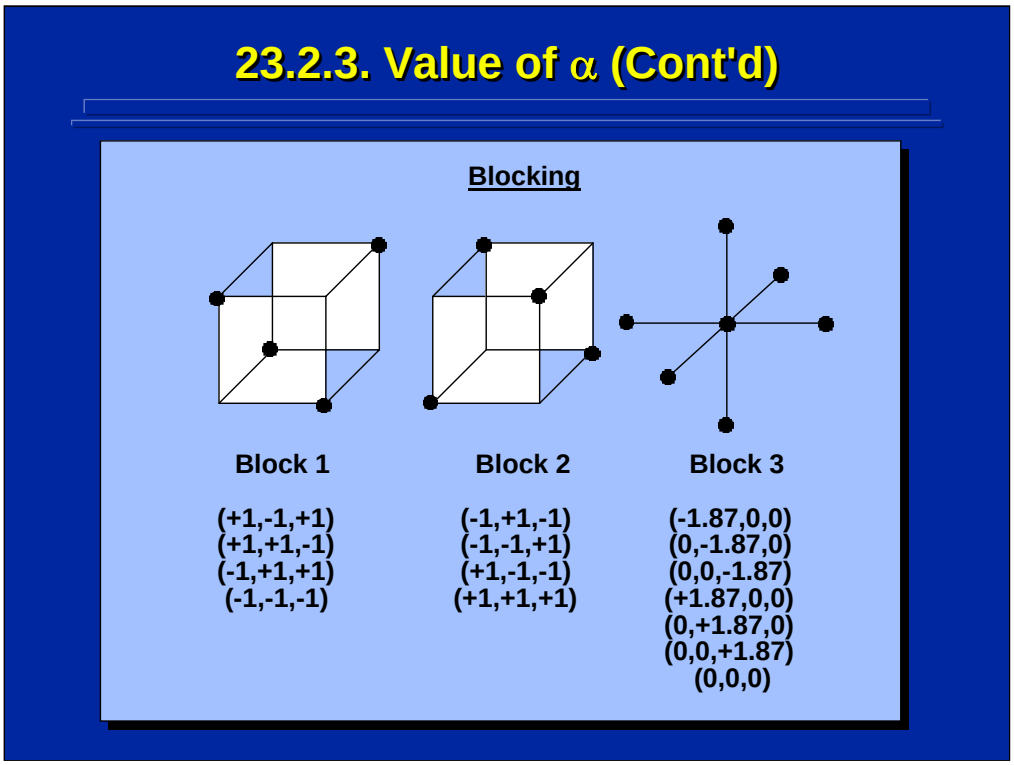

Rotatability, blocking, orthogonal beta weights, spherical, and cuboidal are the five major criteria used in choosing the coded value of $\alpha$ in a CCD. Details on calculating and choosing alternative criteria are discussed by Box and Draper (1987) in chapters 14 and 15 and by Myers and Montgomery (2002) in chapters 7 and 8. Additionally, Williges (1981) describes the calculation of $\alpha$ for behavioral research applications using equal replications across data points in the CCD. A three-factor CCD example is used throughout this subsection for comparing the different coded values of $\alpha$ based on various mathematical criteria.

Rotatability means that the predicted response is the same at all points equidistant from the center. The general equation using rotabability as the mathematical criterion in determining the coded value of $\alpha$ is shown in the middle of this slide and depends on the number of data points in the factorial portion of the CCD. For the three-factor CCD example, $\alpha$ equals 1.682 as shown on the bottom of this slide.

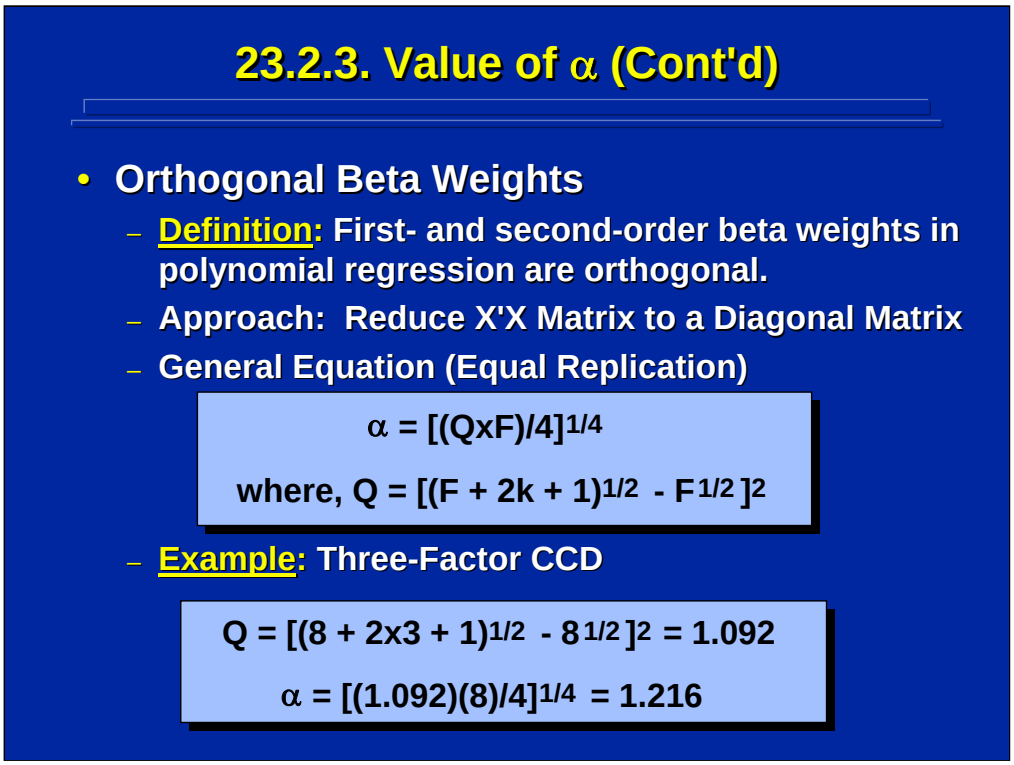



Blocking is used to collect data in stages (e.g., data collection sessions) or control for a nuisance variable (e.g., different experimenters) in the CCD. Block effects are kept orthogonal to the first- and second-order partial regression weights in polynomial regression by keeping the total sum of squares in each block proportional to the number of observations in the block. Usually the factorial portion of the CCD is divided into two blocks using a fractional factorial, and the third block is formed by using the star portion plus the center point of the CCD.

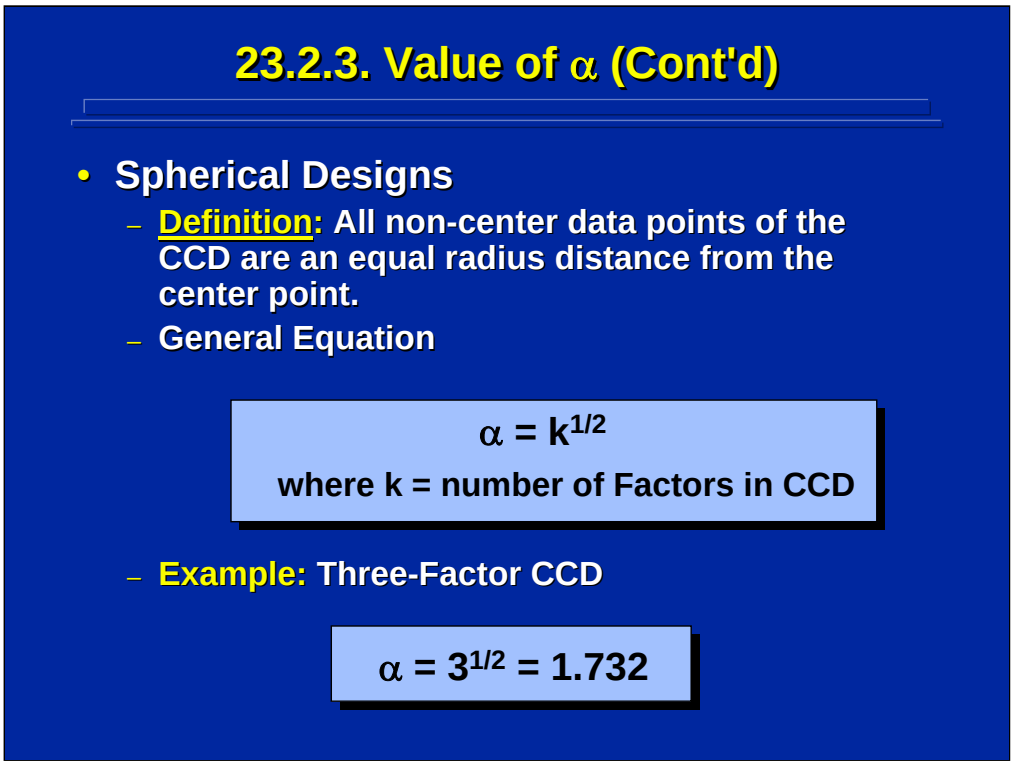
The general formula for $\alpha$ in a blocked CCD based on equal replications across the design is presented in the middle portion of this slide and is simply based on the number of factors, k, in the CCD that determine the number of data points in the third block (i.e., 2k + 1). For the three factor CCD example, $\alpha$ equals 1.871 as shown on the bottom of this slide.



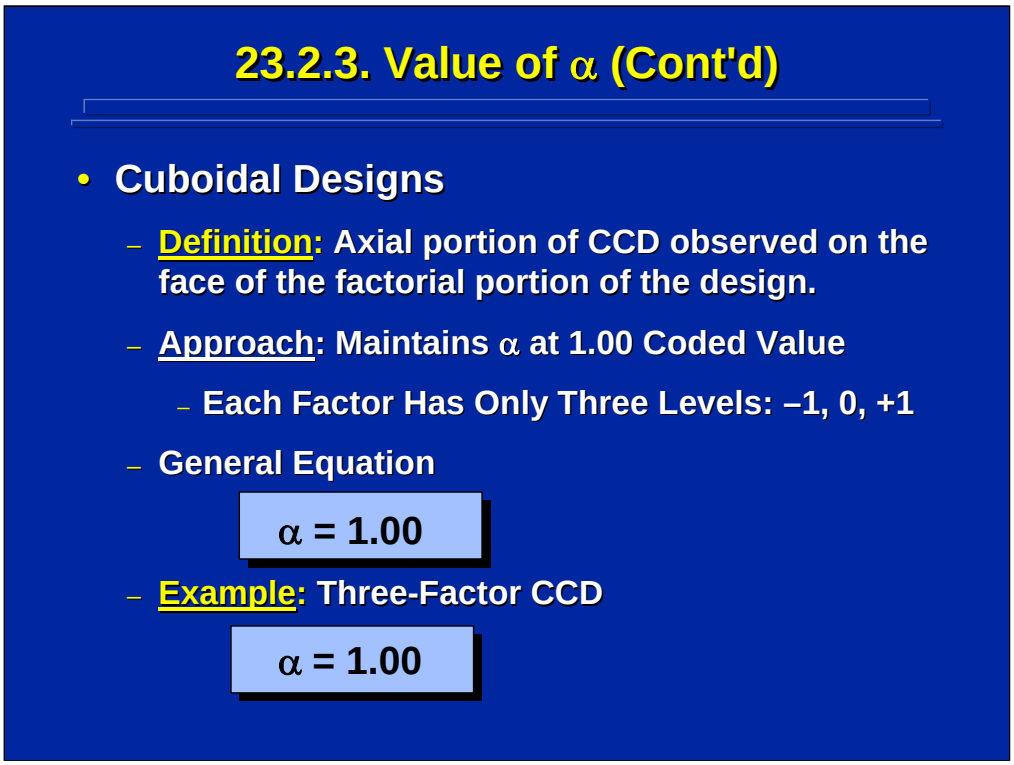

This slide depicts the coded values of the data points for each of the three orthogonal blocks in the three-factor CCD example. Note that the number of data points in the third block has seven treatment combinations (i.e., 2k + 1), whereas the two one-half replicates in the $2^3$ factorial design used in the first two blocks each has four treatment combinations.



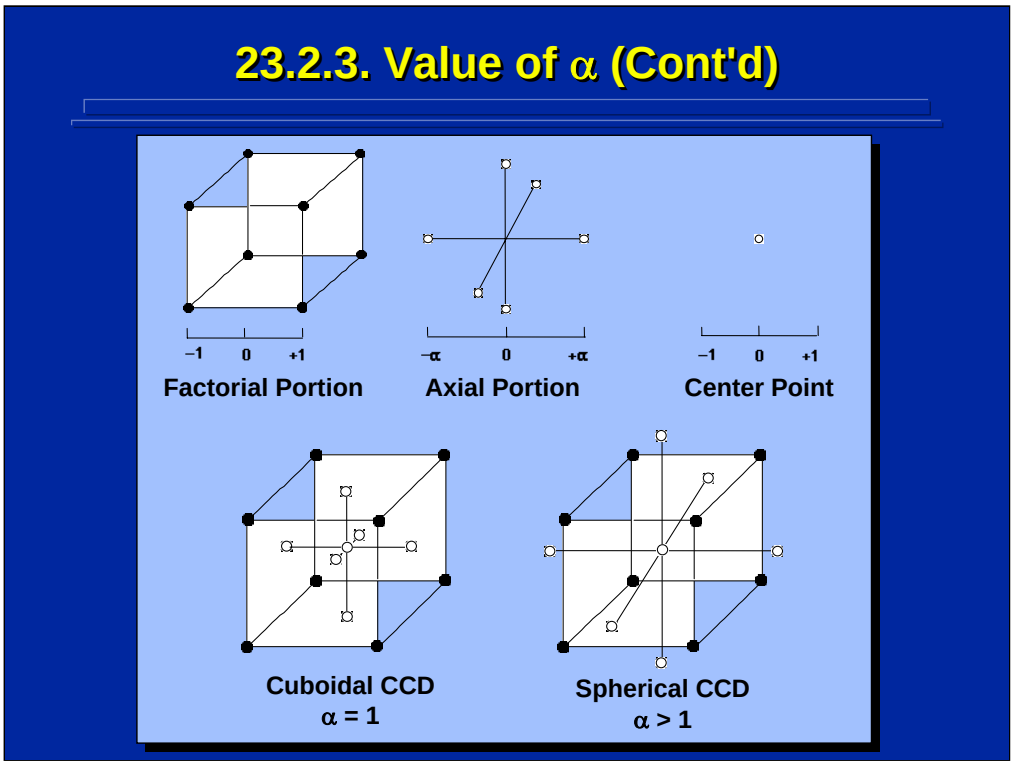

If the experimenter wants to keep all first- and second-order, coded-value partial regression weights orthogonal in the empirical model, then the coded value of $\alpha$ must be adjusted to reduce the X'X matrix to a diagonal matrix when solving the complete second-order polynomial regression model. The general formula for the adjusted $\alpha$ value is presented in the center of this slide when equal replications are used across the CCD. For the three-factor example, an $\alpha$ equal to 1.216 is used in the orthogonal, coded-value CCD as shown on the bottom of this slide.




The rotatable, orthogonal, and blocking alternatives each have the $\alpha$ value extending beyond the ± 1 values of the factorial portion of the CCD, thereby forming a near spherical data collection region. To form a true spherical region, the radius of all non-center points must be equidistance from the center point of the CCD. As shown on this slide this occurs when $\alpha$ is equal to the square root of the number of factors, k, in the CCD. Consequently, $\alpha$ equals 1.732 in the three-factor CCD example as shown on the bottom of this slide.

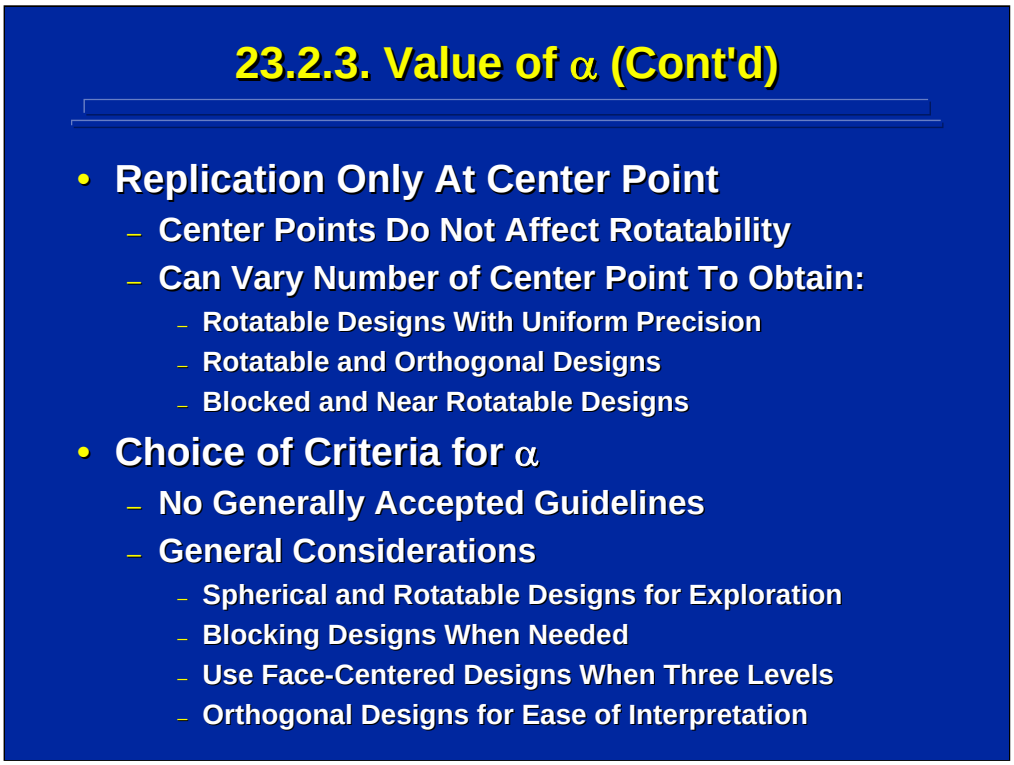



If the region of interest is cuboidal as designated by the $2^k$ factorial portion of the CCD, then the $\alpha$ value can be adjusted by setting the coded value of $\alpha$ equal to 1 so that it appears at the center of each face and does not protrude beyond the face of the $2^k$ factorial portion. A cuboidal CCD is often referred to simply as a “face-centered CCD” and only has three possible coded levels for each factor (i.e., -1, 0, and +1). Consequently, the coded value of $\alpha$ is always 1 regardless of the number of factors in the CCD.




This slide depicts the difference between the cuboidal and spherical design alternatives graphically for a three-factor CCD. The $2^3$ factorial design portion of the CCD forms a cube with the center point in the middle of the cube. When the axial point has a coded value of 1, it appears on the center of each of the six faces of the cube in the cuboidal design, hence a face-centered CCD.

In the spherical CCD alternative, each axial point is greater than 1 and protrudes beyond the center of each face. Consequently, the data points of the $2^3$ factorial portion and 6 axial points fall on the surface, or near the surface, of a ball. When $\alpha$ is greater than 1, the CCD Is designated a spherical, rotatable, orthogonal, or blocked CCD depending on the coded values of $\alpha$.




If the experimenter chooses to replicate only at the center point of the CCD, the number of center points can be varied to provide uniform precision across the rotatable design, orthogonal designs that are rotatable, and blocked designs that are near rotatable. When the CCD has equal replication across the data points, the rotatability criterion is not assured for an orthogonal and blocked CCD. These design alternatives are summarized in Table 1 by Williges (1981, p. 70).

The real choice of $\alpha$ depends on the mathematical criterion the experimenter wishes to emphasize. There are no strict rules for choosing an $\alpha$ criterion and only general guidelines exist as listed on the bottom of this slide. Spherical and rotatable designs are useful when exploring unknown surfaces. For example, Myers and Montgomery (2002, p. 335) recommend that the use of a true spherical design is preferred when the region of interest is spherical. With special experimental design purposes such as staged data collection, existence of a nuisance variable, or the inability to collect five levels of each factor, then alternatives such as a blocking or face-centered CCD needs to be considered. If the resulting second-order empirical model is used to assess the relative importance of the partial regression weights in predicting performance, then an orthogonal CCD is needed and determines the coded value of $\alpha$.

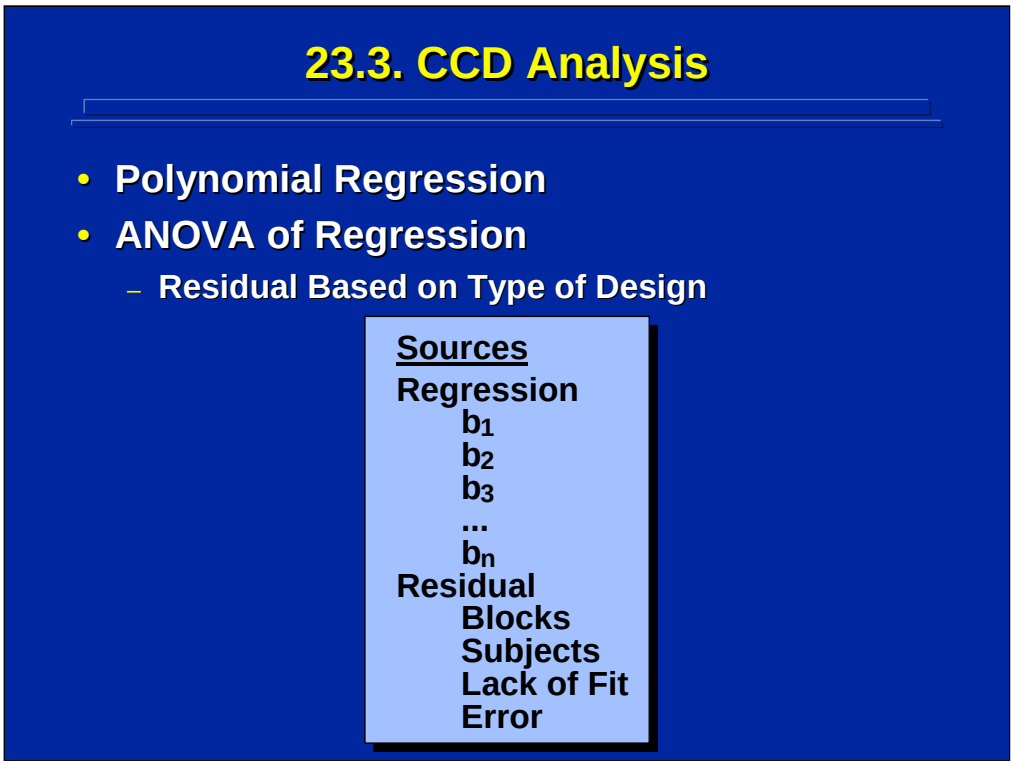
## 23.2.3. Value of α (Cont'd)

### Values of α for Central-Composite Designs With Equal Replication

| Factors (K) | Value of α | | | | | Unique Data Points |
|---|---|---|---|---|---|---|
| | Rotatable | Orthogonal | Blocked | Spherical | Cuboidal | |
| 2 | 1.414 | 1.000 | 1.581 | 1.414 | 1.000 | 9 |
| 3 | 1.681 | 1.216 | 1.871 | 1.732 | 1.000 | 15 |
| 4 | 2.000 | 1.414 | 2.121 | 2.000 | 1.000 | 25 |
| 5* | 2.000 | 1.546 | 2.345 | 2.236 | 1.000 | 27 |
| 6* | 2.378 | 1.724 | 2.550 | 2.449 | 1.000 | 45 |
| 7* | 2.828 | 1.885 | 2.739 | 2.646 | 1.000 | 79 |

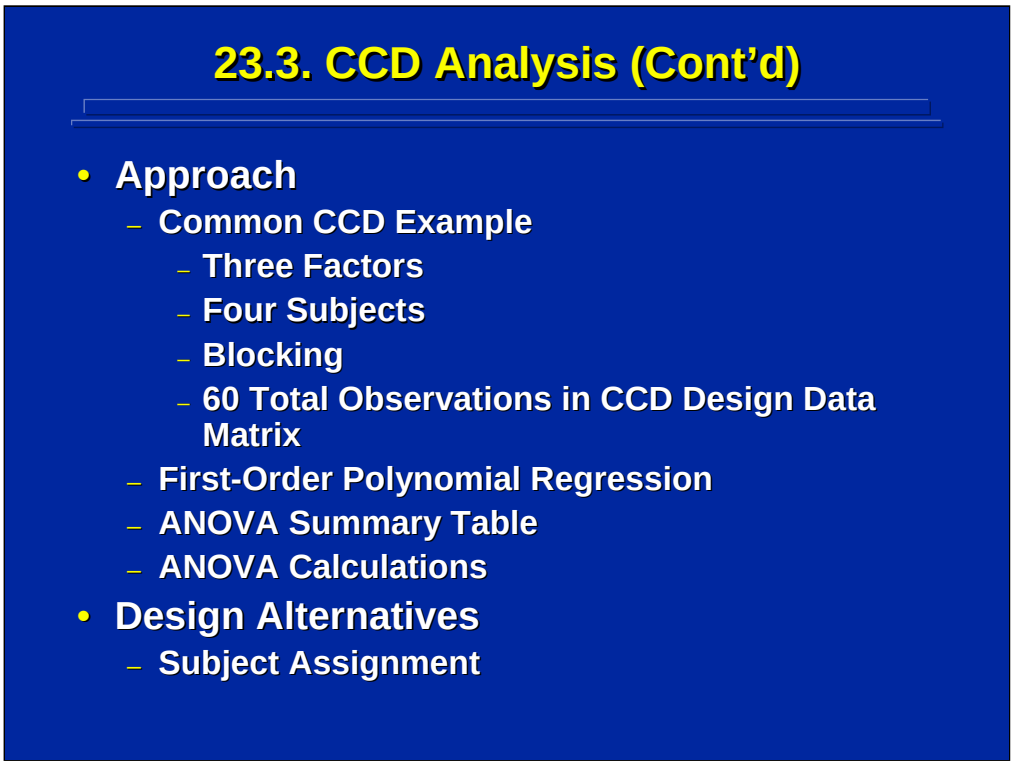
* One-Half Replicate Used in Factorial Portion of Central-Composite Design

This slide summarizes the various coded values of α based on the formulae presented in the previous slides for a two- to seven-factor CCD that has equal replication across the design. The examples provided on previous slides were based on a three-factor CCD. By definition, the coded value of α for a cuboidal CCD is always 1. All the other coded values of α are greater than 1 for any rotatable, orthogonal, blocked, and spherical CCD. Consequently, the coded value of α depends on the specific criterion chosen and the number of factors included in the experimental design. The right most column provides the unique data points used in calculating the various α values presented in the table. Formulae and values of α based on replication only at the center point are provided by Box and Draper (1987) and Myers and Montgomery (2002).

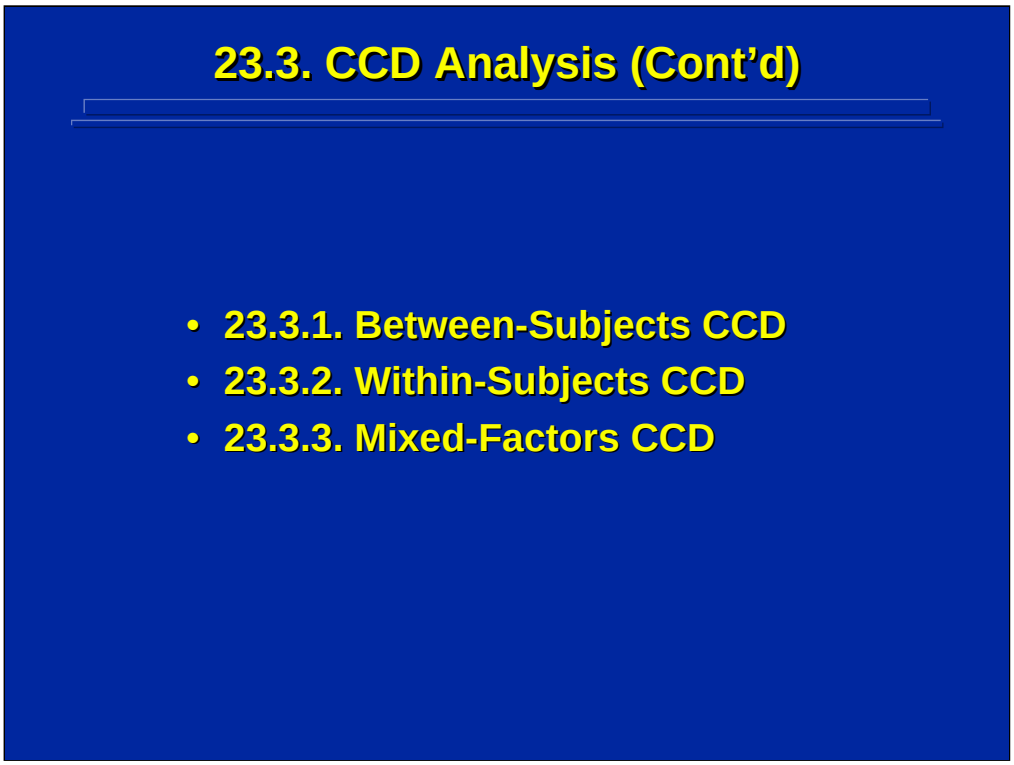



An advantage of the CCD is that the ANOVA of regression is more sensitive than a standard multiple regression analysis of the partial regression weights because the sum of squares of residual error can be decomposed into additive subsets. The possible subsets depend upon the specific CCD used for data collection.

This slide shows the major subsets of residual error that can be determined in a blocked, within-subject CCD in which the block main effect, the subject main effect, lack of fit (LOF), and refined error subsets can be calculated. The LOF effect represents additional partial regression weights that can be used in the regression model. A significant LOF suggests additional partial regression weights can be added to the empirical model that account for a significant amount of variance. The regression partial regression weights, blocks, subjects, and LOF effects can each be tested for significance by using refined error that provides a better estimate of pure error than the pooled residual error term used in standard multiple regression.




A common example of a three-factor, blocked CCD with four observations at each data point yielding a total of 60 observations in the experiment is used to demonstrate the ANOVA on a first-order polynomial regression model. The specific human performance experimental design alternative depends upon the assignment of subjects to treatment conditions.




In this sub-section, between-subjects, within-subjects and mixed-factors CCD alternatives are described using the general three-factor CCD described on the previous slide. Details about these polynomial regression procedures used in behavioral research are presented in Williges (1981, pp. 71-79).

## 23.3.1. Between-Subjects CCD

**Data Matrix**

| Subjects | Treatment | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|---|
| $S_1$ - $S_4$ | 1 | +1 | +1 | +1 |
| $S_5$ - $S_8$ | 2 | -1 | +1 | +1 |
| $S_9$ - $S_{12}$ | 3 | +1 | -1 | +1 |
| $S_{13}$ - $S_{16}$ | 4 | -1 | -1 | +1 |
| $S_{17}$ - $S_{20}$ | 5 | +1 | +1 | -1 |
| $S_{21}$ - $S_{24}$ | 6 | -1 | +1 | -1 |
| $S_{25}$ - $S_{28}$ | 7 | +1 | -1 | -1 |
| $S_{29}$ - $S_{32}$ | 8 | -1 | -1 | -1 |
| $S_{33}$ - $S_{36}$ | 9 | 0 | 0 | 0 |
| $S_{37}$ - $S_{40}$ | 10 | $-\alpha$ | 0 | 0 |
| $S_{41}$ - $S_{44}$ | 11 | $+\alpha$ | 0 | 0 |
| $S_{45}$ - $S_{48}$ | 12 | 0 | $-\alpha$ | 0 |
| $S_{49}$ - $S_{52}$ | 13 | 0 | $+\alpha$ | 0 |
| $S_{53}$ - $S_{56}$ | 14 | 0 | 0 | $-\alpha$ |
| $S_{57}$ - $S_{60}$ | 15 | 0 | 0 | $+\alpha$ |

The fifteen treatment conditions for the three-factor, between-subjects CCD are shown on this slide. Since n = 4 in this design, a different group of four subjects is observed at each treatment combination resulting in a total of 60 different subjects participating in the experiment. Since the CCD is blocked, $\alpha$ equals 1.871 as noted in a previous slide.

## 23.3.1. Between-Subjects CCD (Cont'd)

- **ANOVA Summary Table**

| Source | df | F |
|---|---|---|
| Regression | (3) | |
| $b_1$ | 1 | $MS_{b1}/MS_{Error}$ |
| $b_2$ | 1 | $MS_{b2}/MS_{Error}$ |
| $b_3$ | 1 | $MS_{b3}/MS_{Error}$ |
| Residual | (56) | |
| Blocks | 2 | $MS_{B}/MS_{Error}$ |
| LOF | 9 | $MS_{LOF}/MS_{Error}$ |
| Error | 45 | |
| Total | 59 | |

- **ANOVA Calculations**
  - **Calculate Polynomial Regression**
  - **Refine Residual by Using ANOVA to Determine**
    - **Blocks**
    - **Error (Subjects/Treatments)**
    - **Obtain LOF by Subtraction**

This slide shows the ANOVA summary table for the three-factor, between-subjects CCD shown on the previous slide. There are 3 degrees of freedom for the Regression and 56 for the Residual, combining to a total of 59 degrees of freedom (i.e., 60 total observations minus 1). Since this is a first-order model, only three partial regression weights are included to yield the 3 degrees of freedom for regression.

The 59 degrees of freedom for Residual are separated into 2 for Blocks (i.e. three blocks in the CCD), 9 for Lack of Fit (LOF), and 45 for Replication. The 9 degrees of freedom for LOF are determined by subtracting the degrees of freedom of the first-order model partial regression weights and the blocks effect from the total number of treatments minus 1 (i.e. 15 – 3 – 2 – 1 = 9). The Error effect is the same as subjects nested within treatments in any between-subjects design and its degrees of freedom are determined by the number of treatments times n – 1 (i.e., 15(4 – 1) = 45).

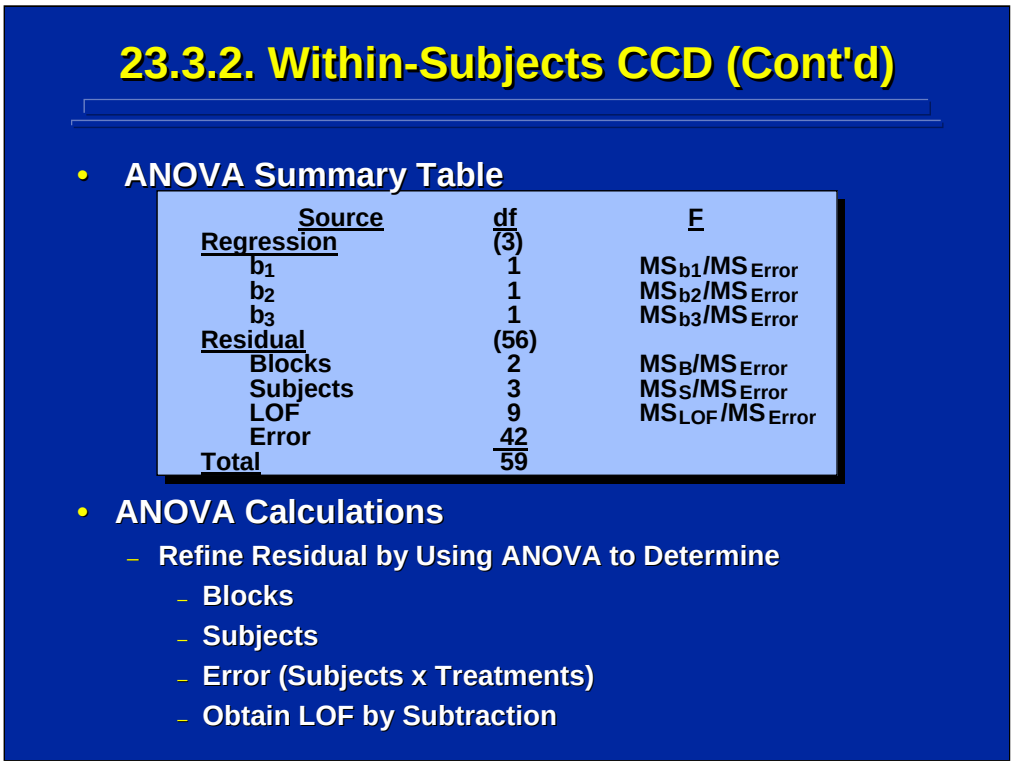
The sum of squares for the main effect of Blocks and Error use standard ANOVA procedures described in Sections 3 and 4 of this reference material, and LOF is obtained by subtraction from total sum of squares. The mean square for Replication can be used as the error term for all F-tests as shown on the slide. Alternatively, the mean square for Residual can be used as a pooled error term for testing the regression model.

## 23.3.2. Within-Subjects CCD

**Data Matrix**

| Subjects | Treatment | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|---|
| $S_1$ - $S_4$ | 1 | +1 | +1 | +1 |
| | 2 | -1 | +1 | +1 |
| | 3 | +1 | -1 | +1 |
| | 4 | -1 | -1 | +1 |
| | 5 | +1 | +1 | -1 |
| | 6 | -1 | +1 | -1 |
| | 7 | +1 | -1 | -1 |
| | 8 | -1 | -1 | -1 |
| | 9 | 0 | 0 | 0 |
| | 10 | $-\alpha$ | 0 | 0 |
| | 11 | $+\alpha$ | 0 | 0 |
| | 12 | 0 | $-\alpha$ | 0 |
| | 13 | 0 | $+\alpha$ | 0 |
| | 14 | 0 | 0 | $-\alpha$ |
| | 15 | 0 | 0 | $+\alpha$ |

The fifteen treatment conditions for the three-factor, within-subjects CCD are shown on this slide. Since n = 4 and subjects are crossed with treatment conditions in a within-subjects design, only a total of four different subjects are needed for this experiment. Each of these four subjects participates in each of the 15 treatment combinations to yield a total of 60 observations in the complete experiment. Again, the CCD is blocked and $\alpha$ equals 1.871 as noted in a previous slide.




This slide depicts the ANOVA summary table for the within-subjects CCD alternative described on the previous slide. Note that Residual now includes the main effect of Subjects. The three degrees of freedom of Subjects is subtracted from Error as compared to the between-subjects CCD alternative. The Error effect is the same as the Subjects x Treatments interaction in any within-subjects design with degrees of freedom equal to (t – 1)(n – 1) or (15 – 1)(4 – 1) = 42.

Calculation of sum of squares, mean squares, and choice of error terms for F-tests on effects are the same as those discussed for the between-subjects CCD alternative. The only changes are the inclusion of the Subjects main effect and the calculation of Error as the Subjects x Treatments interaction.

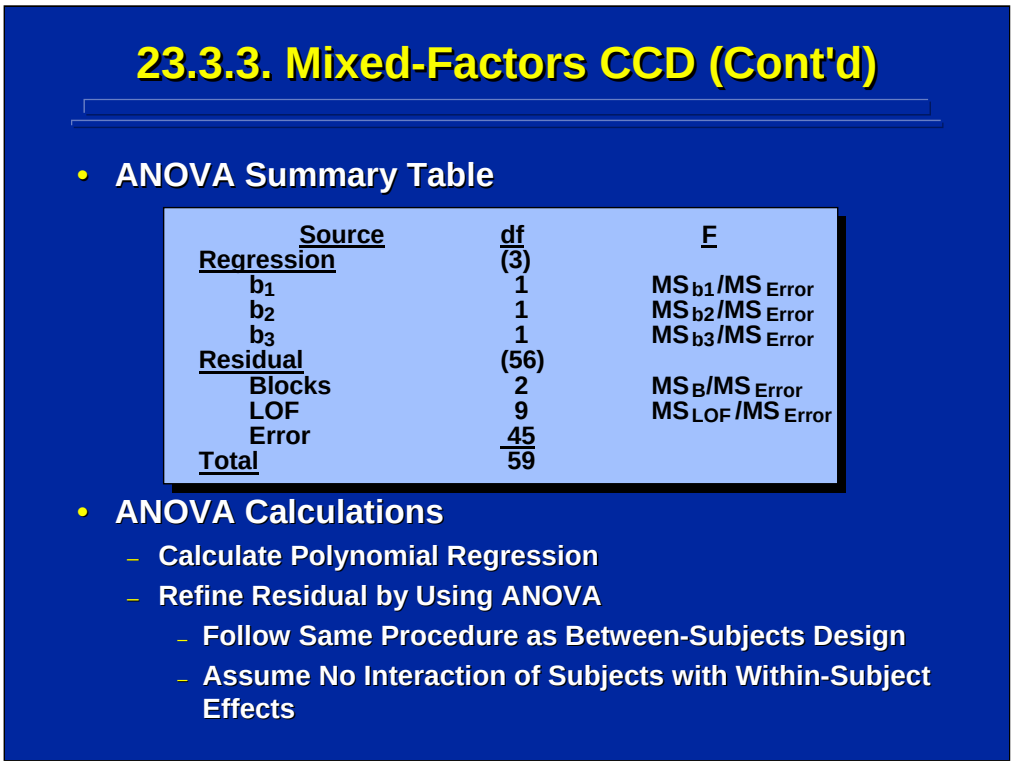
## 23.3.3. Mixed-Factors CCD

Data Matrix

| Subjects | Treatment | Within $X_1$ | Within $X_2$ | Between $X_3$ |
|---|---|---|---|---|
| $S_1$ - $S_4$ | 1 | +1 | +1 | +1 |
| | 2 | -1 | +1 | +1 |
| | 3 | +1 | -1 | +1 |
| | 4 | -1 | -1 | +1 |
| $S_5$ - $S_8$ | 5 | +1 | +1 | -1 |
| | 6 | -1 | +1 | -1 |
| | 7 | +1 | -1 | -1 |
| | 8 | -1 | -1 | -1 |
| $S_9$ - $S_{12}$ | 9 | 0 | 0 | 0 |
| | 10 | $-\alpha$ | 0 | 0 |
| | 11 | $+\alpha$ | 0 | 0 |
| | 12 | 0 | $-\alpha$ | 0 |
| | 13 | 0 | $+\alpha$ | 0 |
| $S_{13}$ - $S_{16}$ | 14 | 0 | 0 | $-\alpha$ |
| $S_{17}$ - $S_{20}$ | 15 | 0 | 0 | $+\alpha$ |

The fifteen treatment conditions for the three-factor, mixed-factors CCD are shown on this slide. Factors 1 and 2 are within-subjects factors and Factor 3 is a between-subjects factor. Consequently, a different group of four subjects is observed at each of the five levels of Factor 3 to provide an n = 4 in each treatment combination.

Since a CCD is not a crossed factorial design, the mixed-factors arrangement results in a differing number of treatment conditions that each subject receives depending on the level of the between-subjects factor the subject is assigned. As shown on this slide, subjects 1 to 8 each receive four treatment combinations, subjects 9 to 12 receive five treatment combinations, and subjects 13 to 20 receive only one treatment combination to result in the 60 observations for the entire CCD experiment. To block the CCD, $\alpha$ equals 1.871 just as in the other two versions of this design.




The regression ANOVA Summary Table is shown on the top of this slide for the mixed-factor CCD described on the previous slide. Since some subjects only receive one treatment combination, no subject main effect can be estimated and the regression ANOVA is the same as that used in the between-subjects CCD alternative summarized in a previous slide. To use Error as the refined error term in F-tests, the experimenter assumes Subjects do not interact with the within-subjects treatments.

Due to unbalanced assignment of treatment conditions in a mixed-factors CCD, this design is used only when absolutely necessary when required by the nature of the factors of interest. If a choice of designs is possible, either a between-subjects or a within-subjects CCD alternative is preferred.

## 23.4. CCD Examples

### Human Factors Research Applications Using CCDs

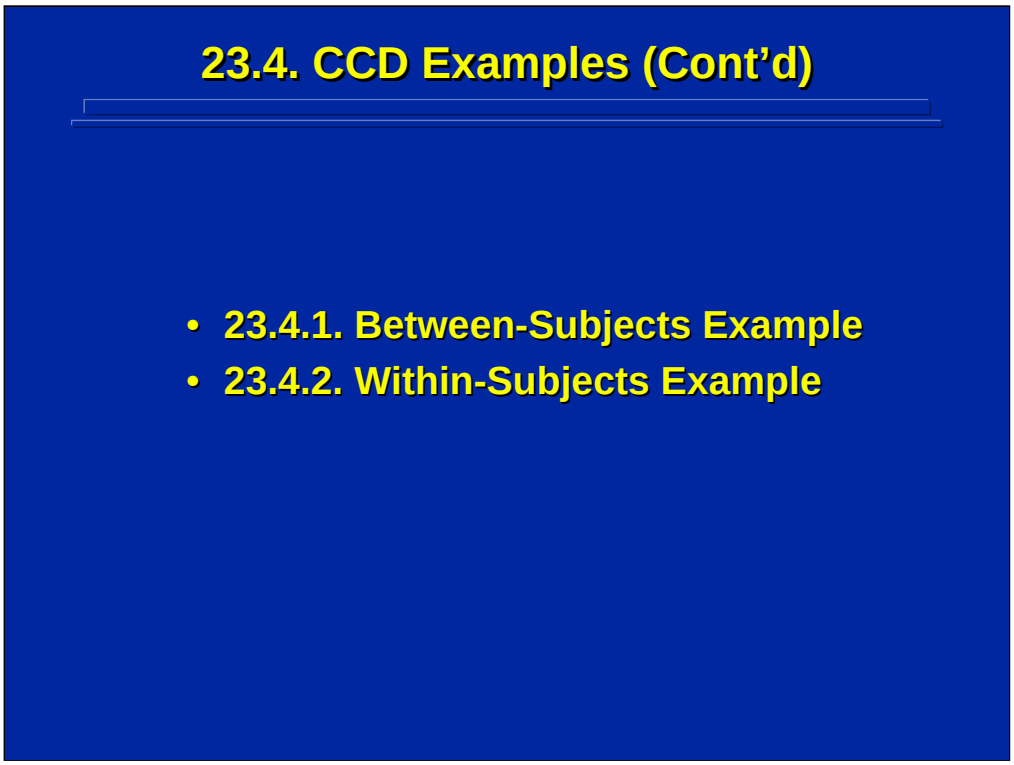
| Reference | CCD α Value | Type of Design | Research Topic |
|---|---|---|---|
| Williges & Baron (1973) | Blocked across Experimenters | Between-Subjects | Transfer of Training (3 Factors) |
| Williges & North (1973) | Blocked across Days | Within-Subjects | Video Cartographic Symbols (4 Factors) |
| Mills & Williges (1973) | Rotatable | Within-Subjects | Surveillance System (5 Factors) |
| *Williges & Williges (1982) | Orthogonal | Between-Subjects | Computer Data Entry (4 Factors) |
| Spine, Williges, & Maynard (1984) | Orthogonal | Within-Subjects | Speech Recognition (4 Factors) |

*Used 22 Dependent Variables

This slide summarizes references to five applications of central-composite designs to human factors research. Across these examples, coded values for $\alpha$ were chosen to provide blocked, rotatable, or orthogonal second-order models using both between-subjects and within-subjects design alternatives. Details on these central-composite design and analysis examples are provided in each reference.

Note that the Williges and Williges (1982) study collected data on 22 dependent variables representing user satisfaction ratings, work sampling procedures, and embedded performance metrics that were subsequently collapsed into three categories using a principal components analysis. Three empirical models were generated based on each of the three multivariate classes of dependent variables representing operator waiting, planning, and production activities.

These early examples were quite successful in generating empirical models of human performance with small sample sizes using a CCD. The resulting polynomial regressions provided high multiple correlation coefficients, R, that were stable under cross validation (Williges and North, 1973) and within $R_{adj}$ prediction (Williges and Mills, 1973).




Two examples of a CCD are provided in this sub-section to summarize the construction and analysis of the CCD. First, a between-subjects CCD is presented followed by its within-subjects CCD alternative.

## 23.4.1. Between-Subjects Example

- **Example Problem: A computer-generated Army surveillance display is tested to predict the effects of three target characteristics on the probability of target detection. The three parameters of interest are target size, target density, and target velocity. Forty-five soldiers were tested in a between-subjects, orthogonal, central-composite design. Is the complete orthogonal, second-order empirical model significant ($p < 0.05$)? Which predictors are significant, and do significant higher-order predictors exist ($p < 0.05$)?**

(Click in this red rectangle to see SAS calculations for this example.)

This example demonstrates the use of an orthogonal, between-subjects, three-factor CCD with equal replications used to generate a complete second-order empirical model predicting the probability of target detection as a function of target size, density, and velocity. Consequently, 45 different soldiers are needed to collect data on the 15 treatment conditions of the three-factor CCD to provide three replications at each treatment combination (i.e., $n = 3$). The problem requires testing the significance ($p < 0.05$) of the overall second-order model, the individual partial regression weights in the model, and the lack of fit of the empirical model. The SAS programs and resulting analyses for this problem are presented in the Slater and Williges (2006) appendix.

## 23.4.1. Between-Subjects Example (Cont'd)

- **Between-Subjects, Orthogonal CCD Data Set**

| Treatments (T) | Coded Values of Target | | | Probability of Detection (P) Where n = 3 |
|---|---|---|---|---|
| | Size (S) | Density (D) | Velocity (V) | |
| 1 | +1 | -1 | +1 | 0.70, 0.82, 0.78 |
| 2 | +1 | +1 | -1 | 0.63, 0.44. 0.52 |
| 3 | -1 | +1 | +1 | 0.65, 0.67, 0.86 |
| 4 | -1 | -1 | -1 | 0.30, 0.45, 0.26 |
| 5 | -1 | +1 | -1 | 0.49, 0.58, 0.47 |
| 6 | -1 | -1 | +1 | 0.48, 0.56, 0.35 |
| 7 | +1 | -1 | -1 | 0.53, 0.74, 0.63 |
| 8 | +1 | +1 | +1 | 0.85, 0.98, 0.81 |
| 9 | -1.216 | 0 | 0 | 0.36, 0.47, 0.55 |
| 10 | 0 | -1.216 | 0 | 0.53, 0.74. 0.60 |
| 11 | 0 | 0 | -1.216 | 0.58, 0.35, 0.25 |
| 12 | +1.216 | 0 | 0 | 0.77, 0.93, 0.81 |
| 13 | 0 | +1.216 | 0 | 0.62, 0.93, 0.68 |
| 14 | 0 | 0 | +1.216 | 0.86, 0.94, 0.96 |
| 15 | 0 | 0 | 0 | 0.75, 0.73, 0.62 |

(Click in this red rectangle to see SAS calculations for this example.)

The slide shows the hypothetical data of the three replications for each of the fifteen data points in the three-factor CCD example problem described on the previous slide. The coded values of the three factors used in the CCD are listed for each of the fifteen treatment conditions. Note that the $\alpha$ values equal ±1.216 in order to determine orthogonal partial regression weights for the second-order model as described in Section 23.2.3.

## 23.4.1. Between-Subjects Example (Cont'd)

**Coded ANOVA Summary Table: Probability of Target Detection (P)**

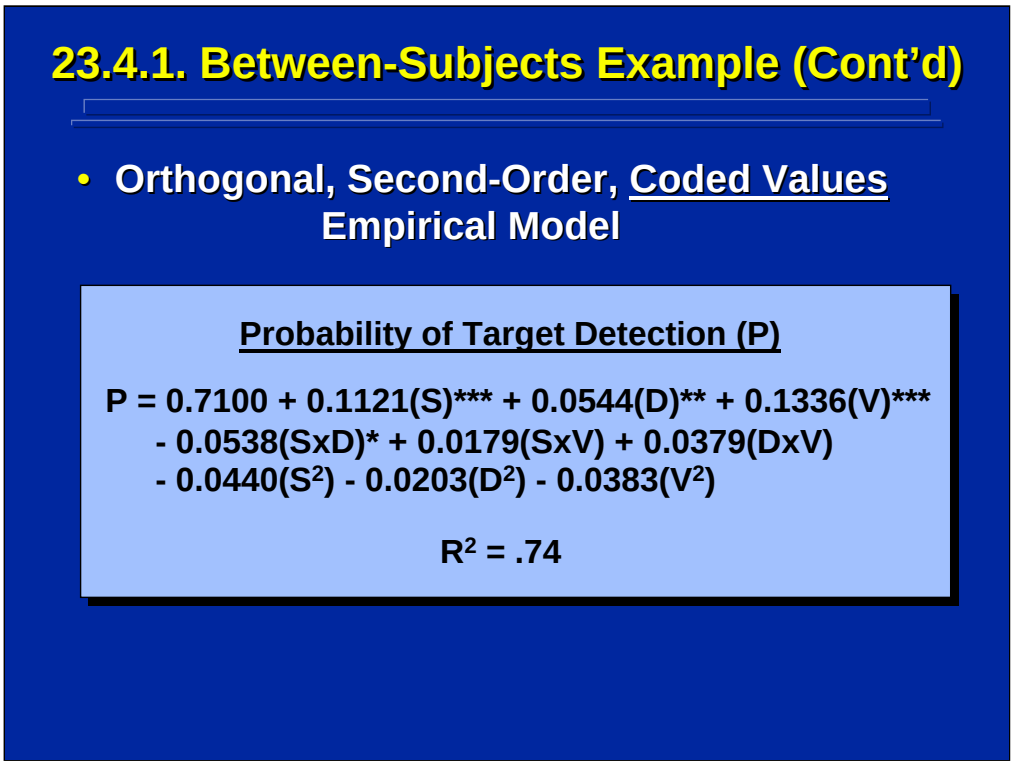
| *Source* | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Regression | (9) | (1.2582) | (0.1398) | (11.26)*** |
| Target Size (S) | 1 | 0.4129 | 0.4129 | 33.25*** |
| Target Density (D) | 1 | 0.0972 | 0.0972 | 7.83** |
| Target Velocity (V) | 1 | 0.5866 | 0.5866 | 47.25*** |
| SxD | 1 | 0.0693 | 0.0693 | 5.58* |
| SxV | 1 | 0.0077 | 0.0077 | 0.62 |
| DxV | 1 | 0.0345 | 0.0345 | 2.78 |
| $S^2$ | 1 | 0.0253 | 0.0253 | 2.03 |
| $D^2$ | 1 | 0.0054 | 0.0054 | 0.43 |
| $V^2$ | 1 | 0.0192 | 0.0192 | 1.55 |
| Residual | (35) | (0.4346) | (0.0124) | |
| Lack of Fit | 5 | 0.1131 | 0.0226 | 2.11 |
| Error | 30 | 0.3215 | 0.0107 | |
| Total | 44 | 1.6927 | | |

*p < 0.05 **p < 0.01 ***p < 0.001

(Click in this red rectangle to see SAS calculations for this example.)

This slide shows the Summary Table of the ANOVA on regression using coded values. There are 9 degrees of freedom for Regression representing the total of the individual partial regression weights for the complete three-factor, second-order model. The sum of squares of the partial regression weights sum to the Regression sum of squares because an orthogonal CCD was used in data collection. The Residual effect can be divided into Lack of Fit and Error. The degrees of freedom for Lack of Fit represent five additional parameters that could be included in the empirical model, and the sum of squares value is determined by subtraction once the Error effect is determined. The degrees of freedom for the Error effect equals t(n – 1), and the sum of squares for Error is the same as the Subjects/Treatments in a between-subjects design.

The F-tests for Regression and the partial regression weights use pooled Residual as the error term. The regression model and four partial regression weights are significant ($p < 0.05$). The F-test on Lack of Fit uses the Error effect as the error term and is not significant at the 0.05 level. Consequently, the second-order model is adequate, and higher-order effects are not required. Alternatively, the Error effect could be used in the denominator for each F-test as a refined estimate of pure error.




(Click in this red rectangle to see SAS calculations for this example.)

This slide lists the complete second-order empirical model predicting detection probability as a function of three target characteristics represented in terms of coded values in the example problem. Based on the regression ANOVA summarized on the previous slide, the three linear effects of target size, density, and velocity as well as the linear-by-linear interaction of target size and density are significant predictors of target detection probability ($p < 0.05$). Since this empirical model is orthogonal, the relative influence of the predictors can be evaluated in terms of the values of the regression weights. For example, target size and velocity each predict twice as much variability as either target density or the linear component of target size-by-density interaction predictors.

The Coefficient of Determination shown at the bottom of this slide states that 74% of the variation in predicting target detection probability is accounted for by the complete second-order model represented by the Regression effect.

## 23.4.1. Between-Subjects Example (Cont'd)

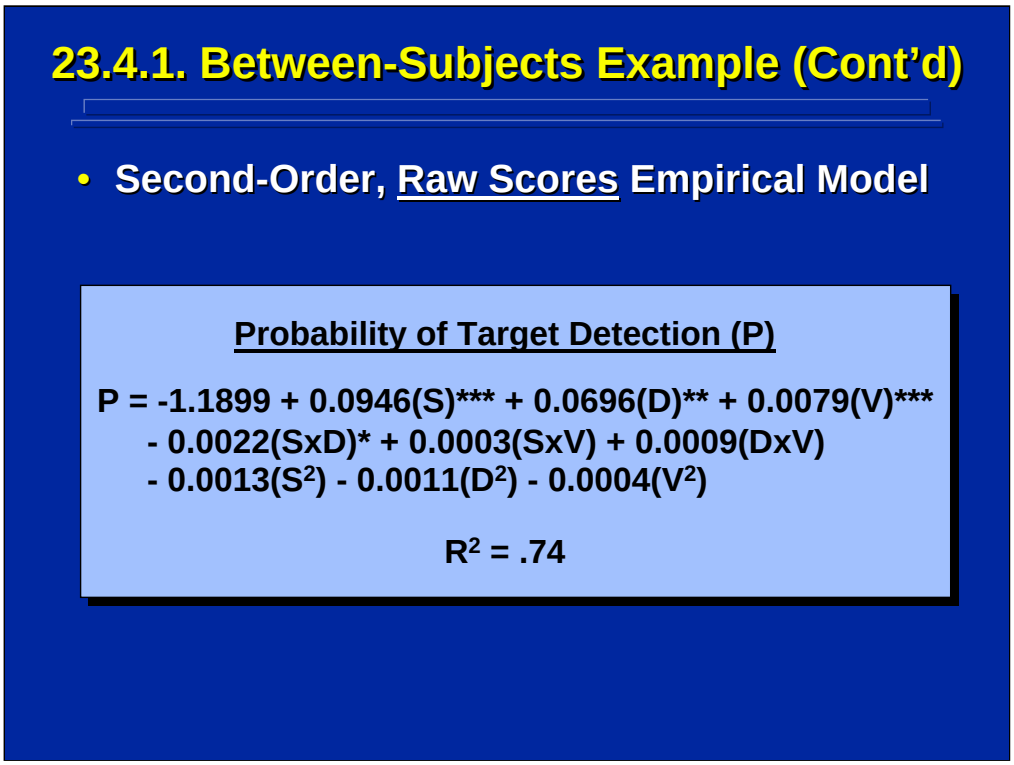
- **Raw Score Linear Transformation**

| | Coded Values | | | | |
|---|---|---|---|---|---|
| | -1.216 | -1 | 0 | +1 | +1.216 |
| Target Size (Height in Pixels) | 11 | 12 | 18 | 24 | 25 |
| Target Density (Targets per Hour) | 11 | 12 | 16 | 20 | 21 |
| Target Velocity (Kilometers per Hour) | 8 | 10 | 20 | 30 | 32 |

(Click in this red rectangle to see SAS calculations for this example.)

The empirical model shown on the previous slide for the example problem is based on coded levels for the three predictors. The experimenter needs to make a linear transformation between these coded levels and real-world levels of the predictors in order to determine an empirical model specified in raw scores. To make this transformation, the experimenter chooses the real world values representing the ±1 coded values and the center of that range equals the 0 coded value. The range of the real world values between the 0 and the ±1 coded values is adjusted by linear transformation to provide the real world values of the $\pm\alpha$ coded values.

This slide summarizes the linear transformation between coded values and raw scores for the five levels of target size, density, and velocity used in the CCD. Target size is specified in terms of height and ranges from 11 to 25 pixels. Target density is specified in terms of the number of targets appearing per hour and ranges from 11 to 21 targets. Target velocity ranges from 8 to 32 kilometers per hour. The experimenter must choose the range of the real-world factor that represents a reasonable operating range of each factor for prediction purposes because the resulting empirical model may not be reliable beyond those ranges.




(Click in this red rectangle to see SAS calculations for this example.)

The second-order empirical model shown on this slide is based on a polynomial regression analysis based on the real-world levels of the three predictors as shown on the linear transformation in the previous slide. The Slater and Williges (2006) appendix provides the details on the SAS analysis of these raw scores.

This raw score model is more meaningful for conducting tradeoff predictions among the three factors of interest because the actual factor levels can be used for each of the predictors in the model. The partial regression weights of these raw score values are not orthogonal due to the varying ranges of raw score values and rounding errors in the linear transformation. Consequently, the coded levels are used to evaluate the relative strength of the partial regression weights, and the raw score model is used for actual predictions.

## 23.4.2. Within-Subjects Example

- **Example Problem: A computer-generated Army surveillance display is tested to predict the effects of three target characteristics on the probability of target detection. The three parameters of interest are target size, target density, and target velocity. Three soldiers were tested in a within-subjects, central-composite design that was blocked across three testing days. Is the complete second-order empirical model significant ($p < 0.05$)? Which predictors are significant and do significant higher-order predictors exist ($p < 0.05$)?**

(Click in this red rectangle to see SAS calculations for this example.)

The second CCD example problem is a within-subjects version of the previous example. In addition, the CCD is blocked across three testing days. Note that only three soldiers are needed in this within-subjects design as compared to the forty-five different soldiers required for the previous between-subjects alternative. The Slater and Williges (2006) appendix provides the SAS solution for this example problem.

## 23.4.2. Within-Subjects Example (Cont'd)

- **Within-Subjects, Blocked CCD Data Set**

| Testing Day (T) | Coded Values of Target Size (S) | Density (D) | Velocity (V) | Probability of Detection (P) Where n = 3 |
|---|---|---|---|---|
| 1 | +1 | -1 | +1 | 0.70, 0.82, 0.78 |
| 1 | +1 | +1 | -1 | 0.63, 0.44. 0.52 |
| 1 | -1 | +1 | +1 | 0.65, 0.67, 0.86 |
| 1 | -1 | -1 | -1 | 0.30, 0.45, 0.26 |
| 2 | -1 | +1 | -1 | 0.49, 0.58, 0.47 |
| 2 | -1 | -1 | +1 | 0.48, 0.56, 0.35 |
| 2 | +1 | -1 | -1 | 0.53, 0.74, 0.63 |
| 2 | +1 | +1 | +1 | 0.85, 0.98, 0.81 |
| 3 | -1.871 | 0 | 0 | 0.36, 0.47, 0.55 |
| 3 | 0 | -1.871 | 0 | 0.53, 0.74. 0.60 |
| 3 | 0 | 0 | -1.871 | 0.58, 0.35, 0.25 |
| 3 | +1.871 | 0 | 0 | 0.77, 0.93, 0.81 |
| 3 | 0 | +1.871 | 0 | 0.62, 0.93, 0.68 |
| 3 | 0 | 0 | +1.871 | 0.86, 0.94, 0.96 |
| 3 | 0 | 0 | 0 | 0.75, 0.73, 0.62 |

(Click in this red rectangle to see SAS calculations for this example.)

This slide lists the coded levels for the within-subjects CCD. Note that the $\alpha$ levels equal 1.871 to provide orthogonal blocking of this three-factor CCD. The left most column of the slide shows the three different data collection days used as blocks. One-half of the treatment conditions in the $2^3$ factorial portion of the CCD were collected on the first day, the other four conditions of the $2^3$ factorial were collected the second day, the remaining seven treatments representing the axial portion and center point of the CCD were collected on day three. The three columns of hypothetical data listed under Probability of Detection represent the data of each of the three soldiers who participated in the experiment.

## 23.4.2. Within-Subjects Example (Cont'd)

**Coded-Values ANOVA Summary Table: Probability of Target Detection (P)**

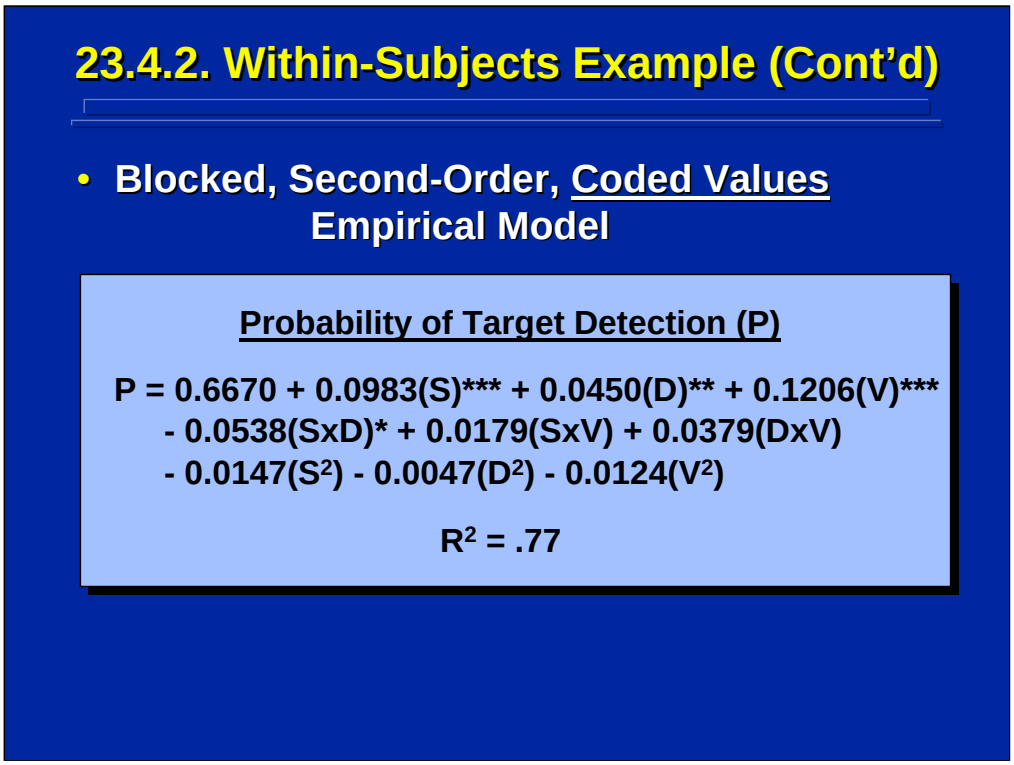
| *Source* | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Regression | (9) | (1.2993) | (0.1444) | (12.84)*** |
| Target Size (S) | 1 | 0.4349 | 0.4349 | 38.69*** |
| Target Density (D) | 1 | 0.0910 | 0.0910 | 8.09** |
| Target Velocity (V) | 1 | 0.6542 | 0.6542 | 58.20*** |
| SxD | 1 | 0.0693 | 0.0693 | 6.17* |
| SxV | 1 | 0.0077 | 0.0077 | 0.69 |
| DxV | 1 | 0.0345 | 0.0345 | 3.07 |
| $S^2$ | 1 | 0.0053 | 0.0053 | 0.47 |
| $D^2$ | 1 | 0.0005 | 0.0005 | 0.05 |
| $V^2$ | 1 | 0.0037 | 0.0037 | 0.33 |
| Residual | (35) | (0.3934) | (0.0112) | |
| Lack of Fit | 3 | 0.0228 | 0.0076 | 0.83 |
| Testing Days | 2 | 0.0492 | 0.0246 | 2.67 |
| Subjects | 2 | 0.0646 | 0.0323 | 3.51* |
| Error | 28 | 0.2568 | 0.0092 | |
| Total | 44 | 1.6927 | | |

*p < 0.05 **p < 0.01 ***p < 0.001

(Click in this red rectangle to see SAS calculations for this example.)

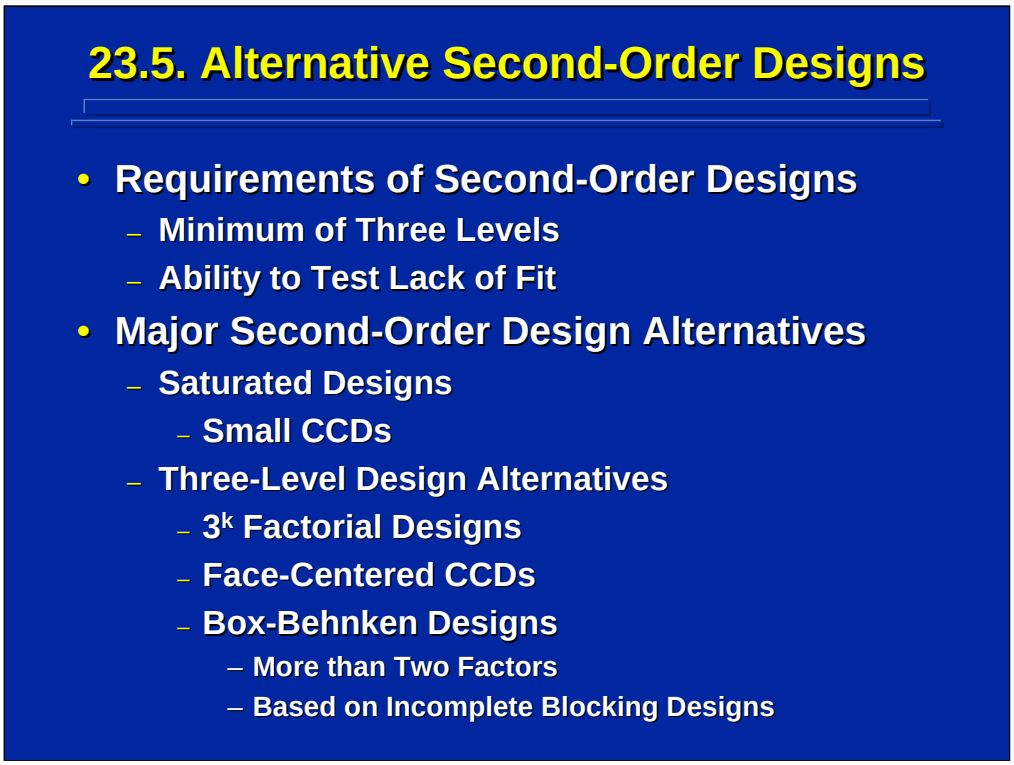
The Summary Table for the ANOVA on regression is shown on this slide. Regression degrees of freedom are the sum of the degrees of freedom associated with the nine partial regression weights in the second-order model, but the sums of squares of these partial regression weights are not orthogonal and do not equal Regression sum of squares. The F-tests on all of these effects are tested by the Residual mean square and show that the regression model and the four partial regression weights each predict a significant amount of probability of target detection variance ($p < 0.05$).

Residual is subdivided into four additive parts with their associated degrees of freedom. The sum of squares of the main effects of Subjects, Testing Days (Blocks), and Error (i.e., Subjects x Treatments interaction) is subtracted from Residual to obtain the sum of squares for Lack of Fit. The F-tests on all four of these effects uses the Error mean square as the denominator in the F-ratio. Since there is a significant Subjects effect ($p < 0.05$), the experimenter could choose not to pool effects and use Error rather than Residual as the error term in all F-tests.




(Click in this red rectangle to see SAS calculations for this example.)

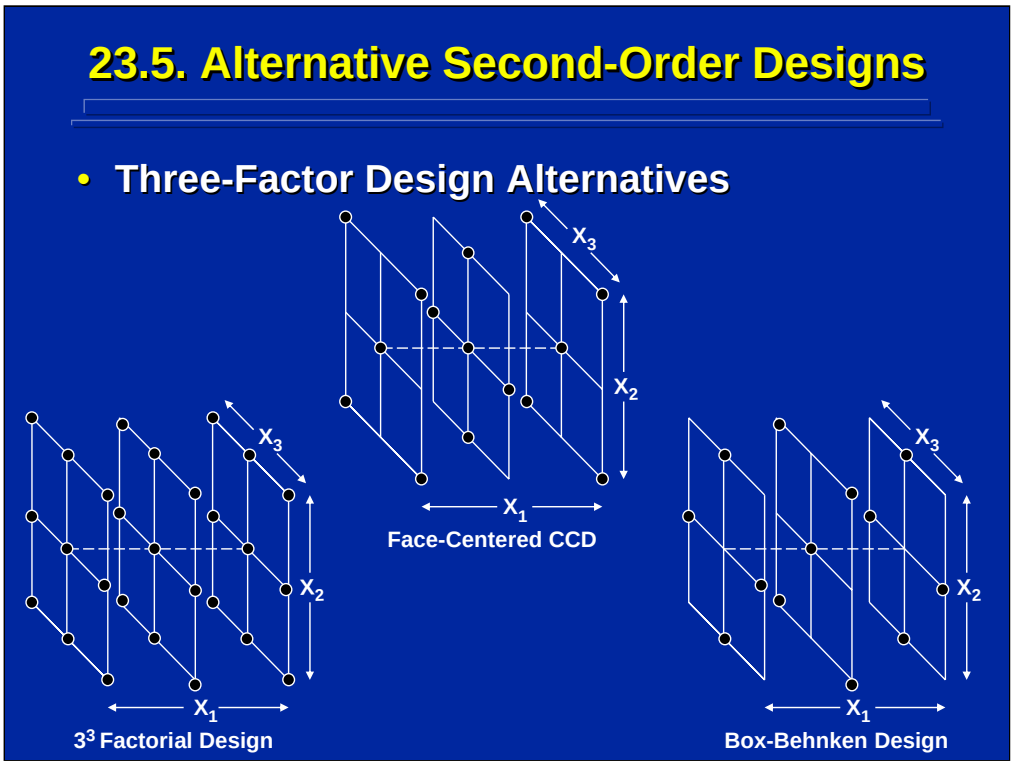
The resulting coded-value, complete second-order empirical for the within-subjects CCD example is shown on this slide. Since this model is not based on an orthogonal second-order design, the values of the partial regression weights differ slightly from the previous between-subjects CCD example. In both cases, however, the regression model and the same four partial regression weights are significant ($p < 0.05$) and the Coefficient of Determination is the same. Raw score levels of the three target parameters as defined by linear transformations of the coded values should be used to generate a raw score empirical model for prediction purposes.




Although the CCD is the primary design of choice in solving second-order empirical models, there are alternatives. These alternative designs must include a minimum of three levels of each factor and provide the capability to test lack of fit for the possibility of higher-order effects.

Two general types of second-order design alternatives exist. A saturated design such as a small CCD requires a relatively small number of runs and is described by Box and Draper (1987, pp. 520-522) and Myers and Montgomery (2002, pp. 378-384). Although these designs minimize the number of runs required to solve second-order models, they often require five levels of each factor as used in a standard CCD.

Three major design alternatives are considered when only three levels of each factor are investigated. These alternatives are $3^k$ factorial designs, face-centered CCDs, and Box-Behnken designs as developed by Box and Behnken (1960). The Box-Behnken designs require a minimum of three factors and are based on incomplete blocking of either a $2^2$ or $2^3$ factorial structure of the design. See Myers and Montgomery (2002, pp. 343-350) for a detailed description of Box-Behnken design alternatives.




This slide illustrates a comparison of data points in a three-factor design for the face-centered CCD, the $3^k$ factorial design, and the Box-Behnken design alternatives. All three observe only three levels of each factor. Unique data points are depicted as small circles on the slide. The face-centered CCD has 15 data points, the $3^3$ factorial design has 27 data points, and the Box-Behnken has 13 data points. Note that the distribution of data points in the face-centered CCD is cubical and the distribution of the data points in the Box-Behnken design is spherical.

## 23.5. Alternative Second-Order Designs

- **Comparison of Second-Order Designs**

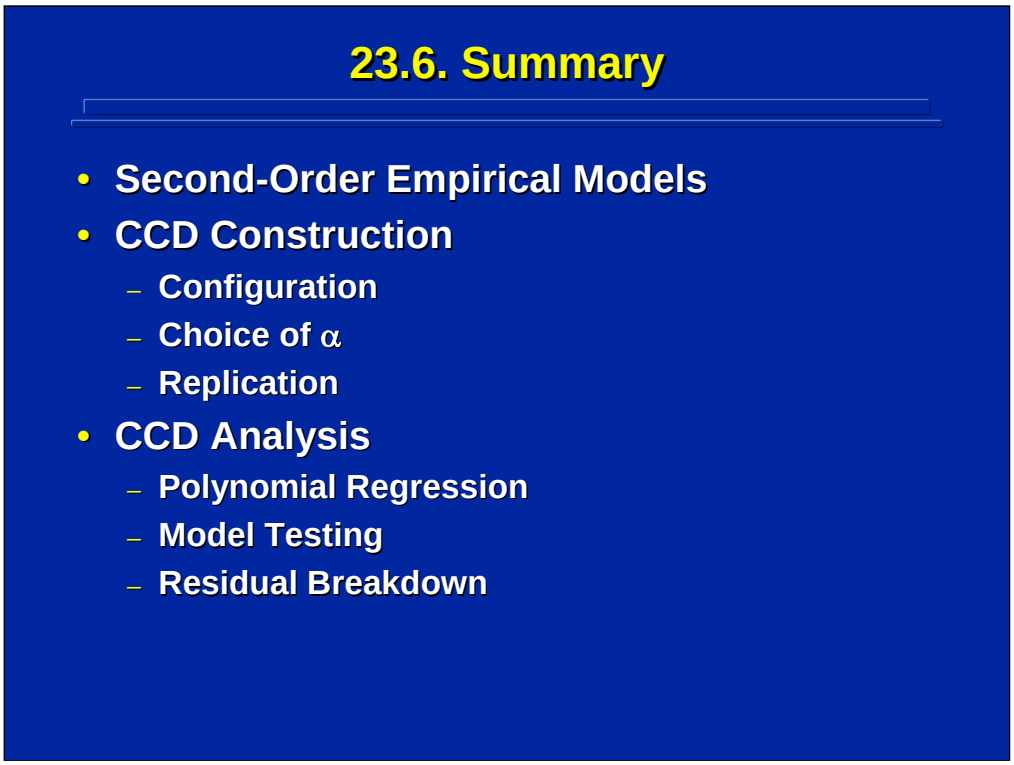
| Number Of Factors | $3^k$ Factorial Designs | Central-Composite Designs | Box-Behnken Designs |
|---|---|---|---|
| 2 | 9 | 9 | --- |
| 3 | 27 | 15 | 13 |
| 4 | 81 | 25 | 25 |
| 5 | 243 | 27* | 41 |
| 6 | 729 | 45* | 49** |
| 7 | 2187 | 79* | 57** |

* Using a One-Half Replicate in the $2^k$ Factorial Portion
** Using $2^3$ Factorial Structure for Incomplete Blocks Portion

- **Advantages of CCD Designs**
  - Expandable to Five Levels of Each Factor
  - Amenable to Sequential Experimentation

This slide compares the number of unique data points in the major three-level experimental design alternatives for generating second-order models based on two through seven quantitative factors. Note that the $3^k$ factorial design alternative quickly becomes uneconomical since it primarily provides data to evaluate effects greater than second order. The CCD and the Box-Behnken design alternatives are somewhat comparable in terms of minimum data points required with the CCD requiring fewer data points for investigating five factors but more data points for investigating seven factors.

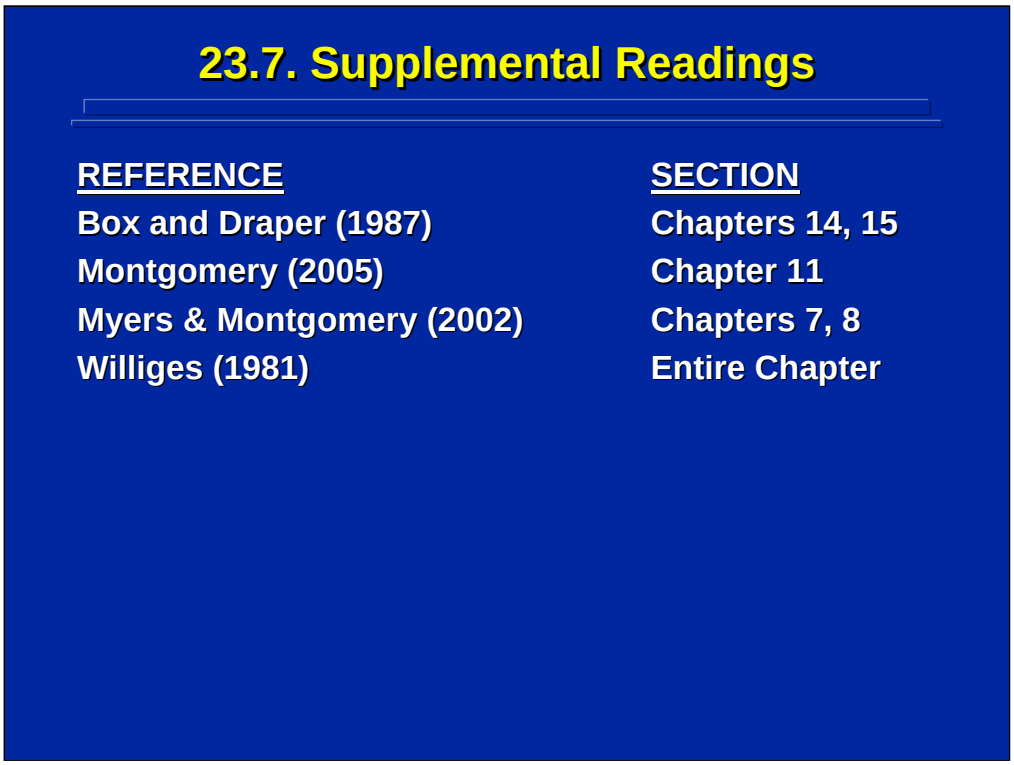
In general, the CCD is most often used to generate second-order empirical models because it can easily expand to investigating five levels of each factor without increasing the number of data points required and the blocking versions of the CCD are readily amenable to sequential experimentation (Williges, 2006). Consequently, this reference material is restricted to a discussion of using the CCD for collecting data to generate second-order empirical models.




By way of summary, this topic describes the use of the CCD as the experimental design for collecting data to solve second-order models of human performance in complex systems. Due to economy of data collection and flexibility in design configuration, the CCD is the experimental design of choice for empirical model building.

The CCD is constructed as a composite of a $2^k$ or $2^{k-p}$ factorial portion with an axial portion and a center point. The coded values of the axial points are defined as $\pm\alpha$ in coded form. The exact value of $\alpha$ depends on mathematical criteria to define a rotatable, orthogonal, blocked, spherical, or faced-centered CCD. Usually replication occurs only at the center point to increase data collection economy. The CCD can be used as between-subjects, within-subjects, or mixed-factors design with equal replication across data points for collecting data in human factors and ergonomics research.

The primary analysis of a CCD is a polynomial regression to represent complete second-order empirical models. A subsequent ANOVA can be conducted on the polynomial regression to test the significance of the regression model and the individual partial regression weights using residual as error. Depending upon the specific CCD used, residual can be separated into subjects, blocks, lack of fit, and error components.




The chapters by Box and Draper (1987), Montgomery (2005), and Myers and Montgomery (2002) provide a general overview of the CCD as well as detailed discussions of design construction. The Williges (1981) article provides details on the construction and use of the CCD in behavioral research.

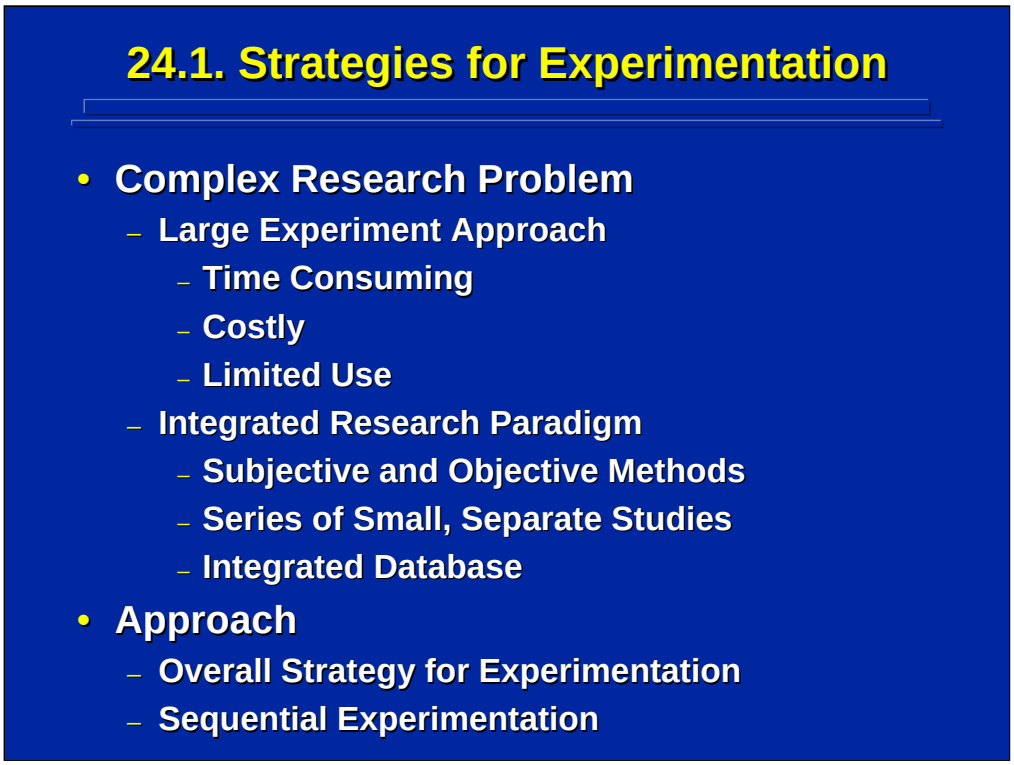
# Topic 24. Sequential Experimentation

24.1. Strategies for Experimentation
24.2. Response Surface Methodology (RSM)
24.2.1. Steps in RSM
24.2.2. Method of Steepest Ascent
24.3. Sequential Research
24.3.1. Sequential Research Paradigm
24.3.2. Sequential Research Example
24.3.3. Guidelines for Sequential Research
24.4. Integrated Research Database
24.5. Summary
24.6. Supplemental Readings

Topic 24 incorporates the previous topics into an overall strategy for conducting research on complex systems often addressed in human factors and ergonomics research. Due to the nature of complex research, the experimenter can conduct a series of small interrelated experiments rather than one large complex study.

This topic describes general strategies for conducting complex experiments and specifically draws upon considerations made in response surface methodology for conducting a series of small, interrelated studies. A general paradigm for conducting sequential research is presented along with a detailed example of using this paradigm in human factors research. This research resulted in a set of guidelines for sequential experimentation.

The topic concludes with a description of combining sequential experiments into a common database that incorporates the results of several experiments. This database can be interrogated to generate integrated empirical models across experiments. The techniques of sequential research are summarized at the end of this topic along with supplemental reading references for details on these procedures.




Many human factors research problems exist in complex systems where human performance is affected by a large number of independent variables that would require a large experimental design. Using one large experiment can quickly become quite time consuming and costly resulting in unwieldy data collection. In addition, the design is of limited value if the researcher is primarily interested in investigating only first-order and second-order effects.

Alternatively, an integrated research paradigm can be chosen that uses both objective and subjective methods to select the independent variables of interest and investigate this subset through a series of small studies. The results can be combined into an integrated database. This topic describes approaches taken to develop an integrated research procedure that results in sequential experimentation in contrast to conducting one large experiment.

Simon (1977a, b) was one of the first human factors researchers to address methods for conducting complex experimentation. As shown on this slide, he recommended five major phases of complex experimentation. Key to his approach was the use of small screening experiments that emphasize investigation of first- and second-order effects with a minimum amount of replication in each experiment.

## 24.1. Strategies for Experimentation (Cont'd)

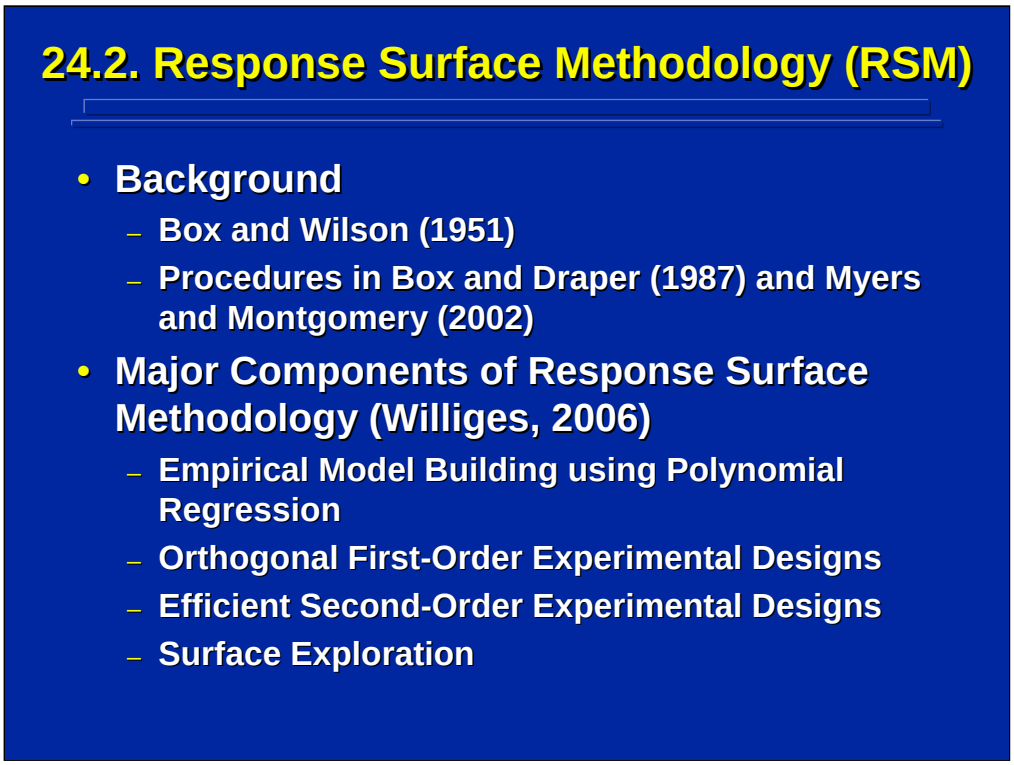
- Mills (1979) Strategy For Experimentation
  - Define Experimental Space
    - Dependent Variables: Performance measures
    - Independent Variables: Controlled conditions occurring in one particular experiment
    - Constants: Experimental conditions held constant across a series of experiments
    - Parameters: Controlled conditions occurring in every experiment in a series
- Experimental Designs
  - Small $2^k$ Factorial Designs
  - Central-Composite Designs

Mills (1979), another human factors researcher, emphasized careful definition of the research space before beginning complex experiments. As shown on this slide, his listing of dependent variables, independent variables and constants follow standard experimental design procedures. But, his inclusion of parameters is an important additional consideration. As stated on this slide, a parameter is an independent variable that is so central to the research problem that it is included in each experiment in a series of small $2^k$ factorial designs and central-composite designs used in sequential experimentation.

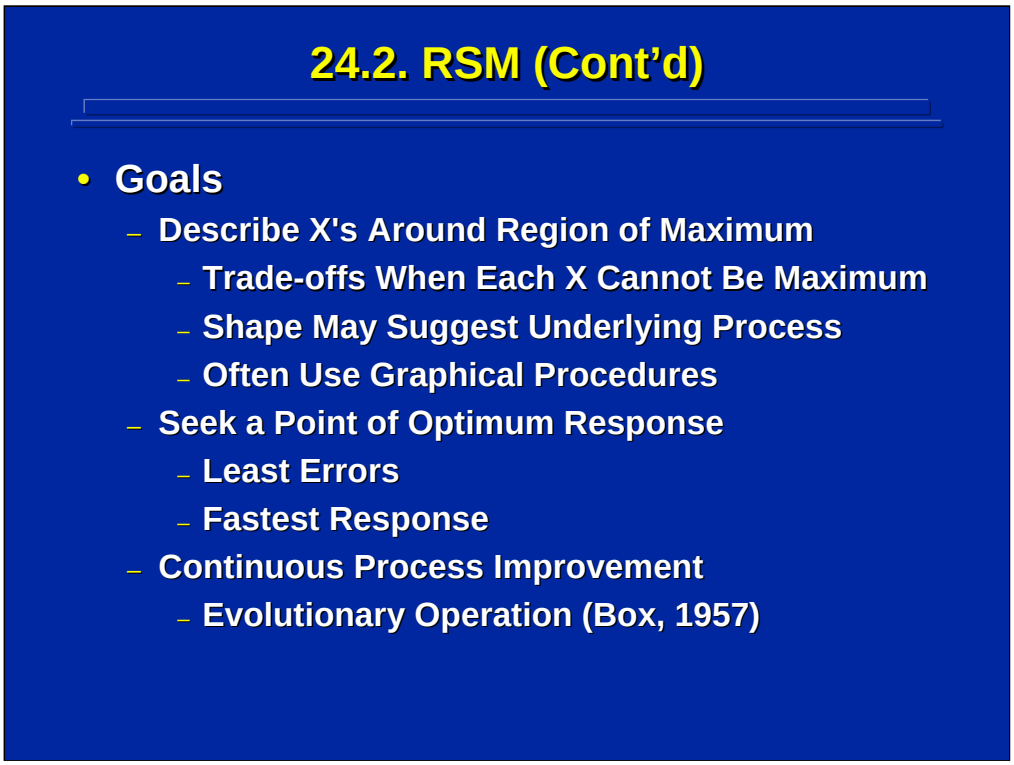
## 24.1. Strategies for Experimentation (Cont'd)

- **Diamond (1981) Strategy For Experimentation**
  - **Overall Strategy**
    - **Define Specific Objectives of Study**
    - **Define Total Experimental Space**
    - **Define Responses of Interest**
  - **Initial Experiments: Specify Significant Variables**
    - **Independent Variables and Range of Interest**
  - **Subsequent Experiments: Estimate Relationships**
    - **Two-Level Experiments**
  - **Final Experiments: Estimate curvature and maxima**
    - **Multilevel Experiments**

The Diamond (1981) textbook recommended an overall strategy of complex experimentation that included defining the research objective, the total experimental space, and the range of response interest. Again, his approach uses a series of sequential experiments incorporated into the three phases shown on this slide. He emphasized the use of $2^{k-p}$ fractional-factorial designs as a means of minimizing treatment conditions when estimating relationships. Multilevel designs were suggested only for the final experiments.

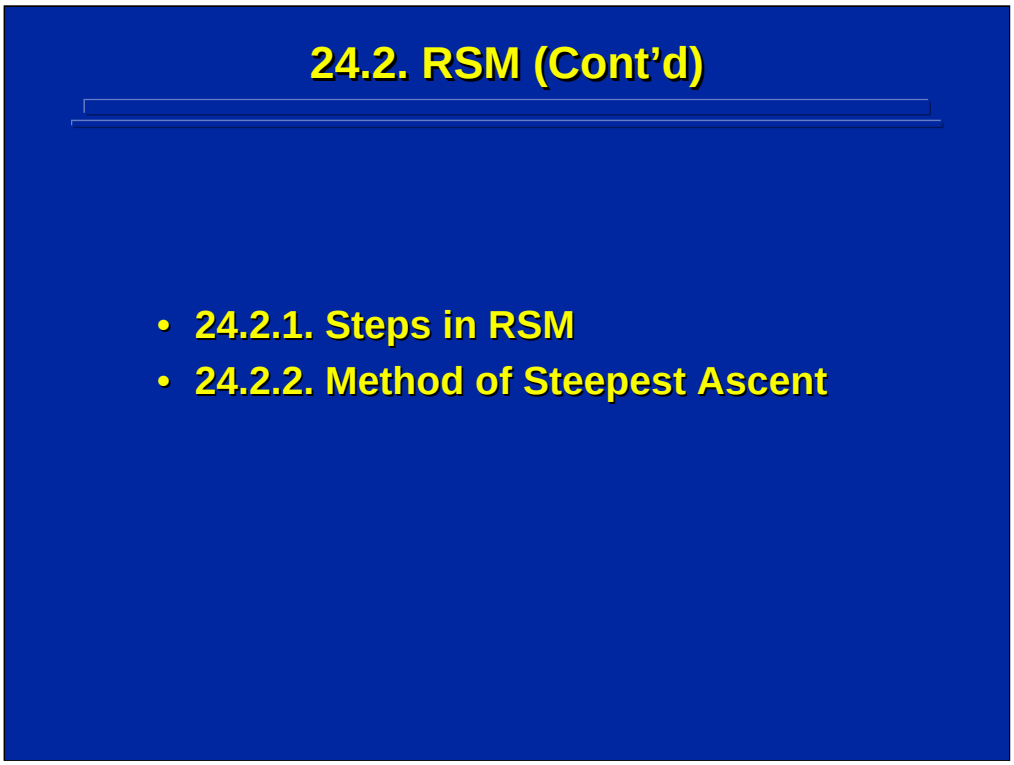



Probably the first comprehensive approach to building empirical models through sequential experimentation was addressed by Box and Wilson (1951) in their discussion of response surface methodology (RSM). These procedures were originally developed for industrial process control to seek optimum yield of chemical reactions. Textbooks by both Box and Draper (1987) and Myers and Montgomery (2002) provide the details of RSM. Williges (2006) summarized the four major components of RSM that are particularly useful to human factors and ergonomics research as listed on the bottom of this slide.

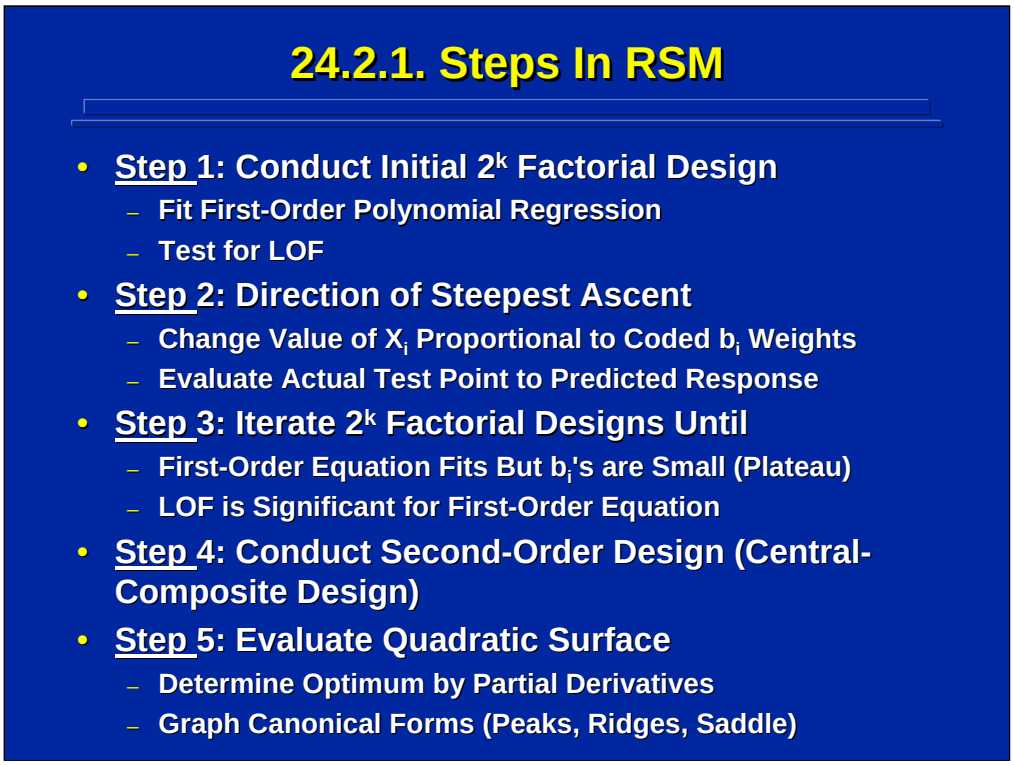



The major goals of RSM are summarized on this slide. Statistical and plotting techniques in RSM are used to describe the region of the response surface around the optimum and to find a point of optimum response, if it exists. In human factors applications, an optimum response is defined using human performance metrics such as the least number of errors or the fastest response time.

One extension of these RSM goals is to use RSM procedures for continuous process improvement through evolutionary operation (EVOP) developed by Box (1957). Myers and Montgomery (2002) in Chapter 14 provide details and examples of EVOP related procedures.

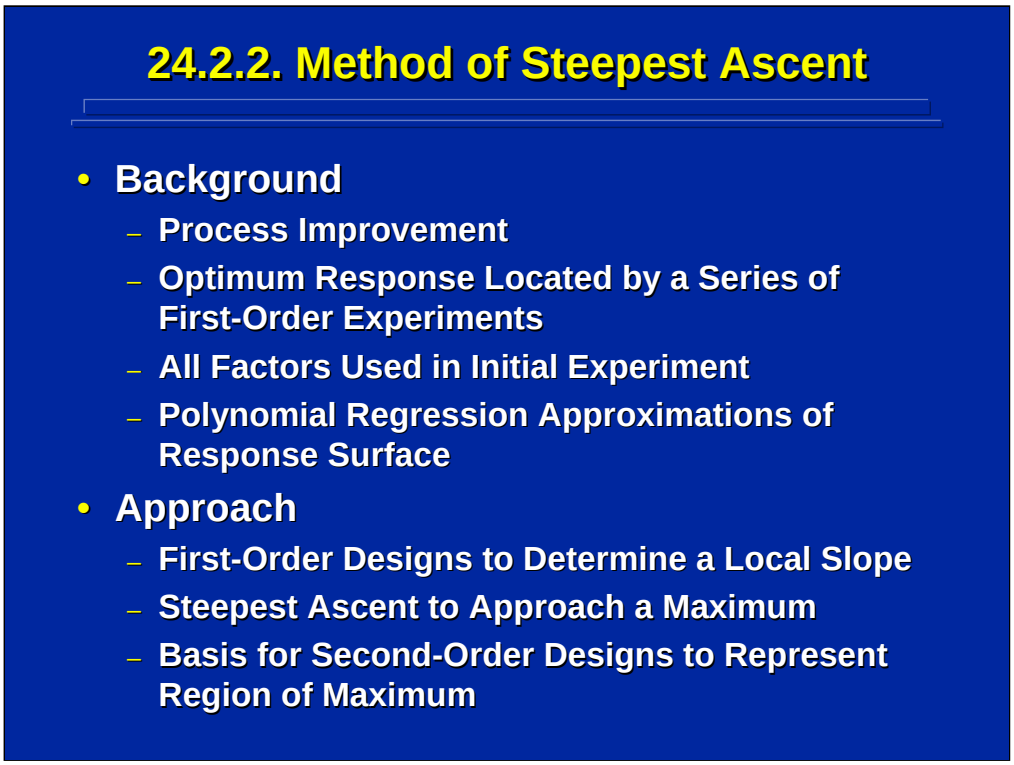



This subsection describes the general steps in RSM and provides details on one particular RSM technique, the method of steepest ascent, which illustrates the sequential data collection philosophy of RSM.



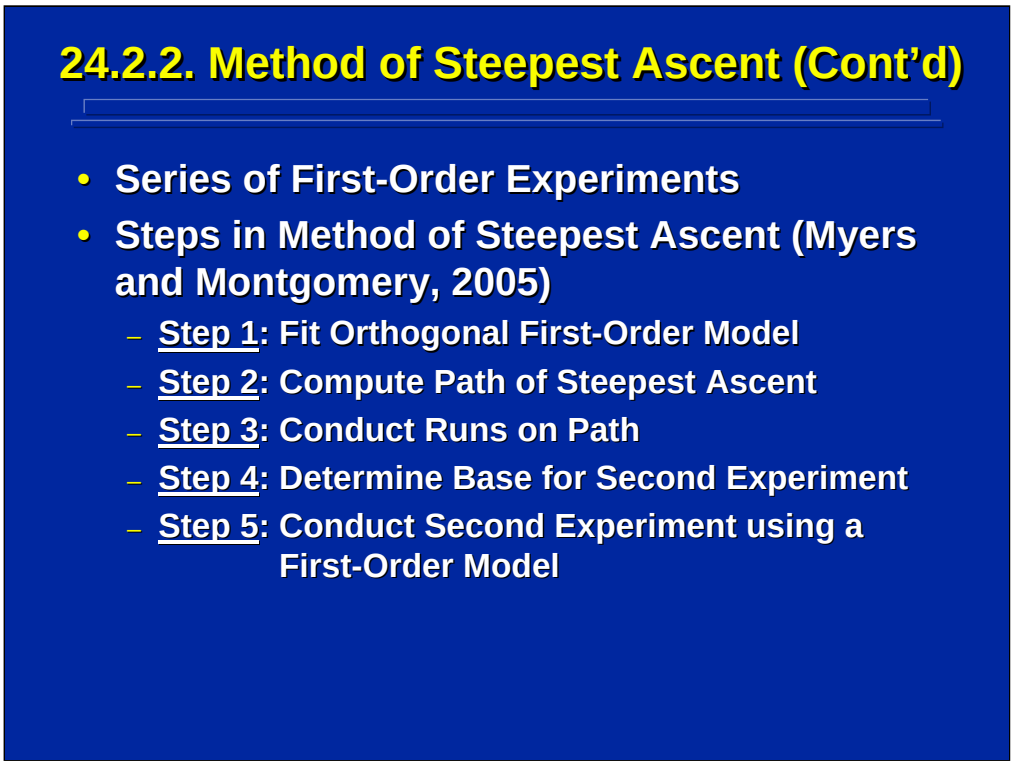

Although RSM is a compilation of several techniques, the overall approach involves sequential data collection through a series of experiments. See Chapters 9 through 12 in Box and Draper (1987), Chapters 11 and 12 in Box, Hunter, and Hunter (2005), Chapter 11 in Montgomery (2005), and Chapter 6 in Myers and Montgomery (2002) for details on RSM procedures.

This slide summarizes the five major sequential steps in RSM. Data collection begins with first-order designs to investigate the influence of major factors affecting the response surface. The method of steepest ascent is used to explore the slopes on major factors rapidly to approach an optimum. First-order designs are used until there is a significant lack of fit suggesting the need for second-order effects. Central-composite designs are the major second-order designs used in RSM to describe the region of optimum performance. The final step in RSM is the evaluation of the region of optimum performance through analytical and graphing procedures. These procedures are used to isolate an optimum point, a series of optimum points (i.e., ridge), or a plane of optimality (i.e., saddle) on the response surface.

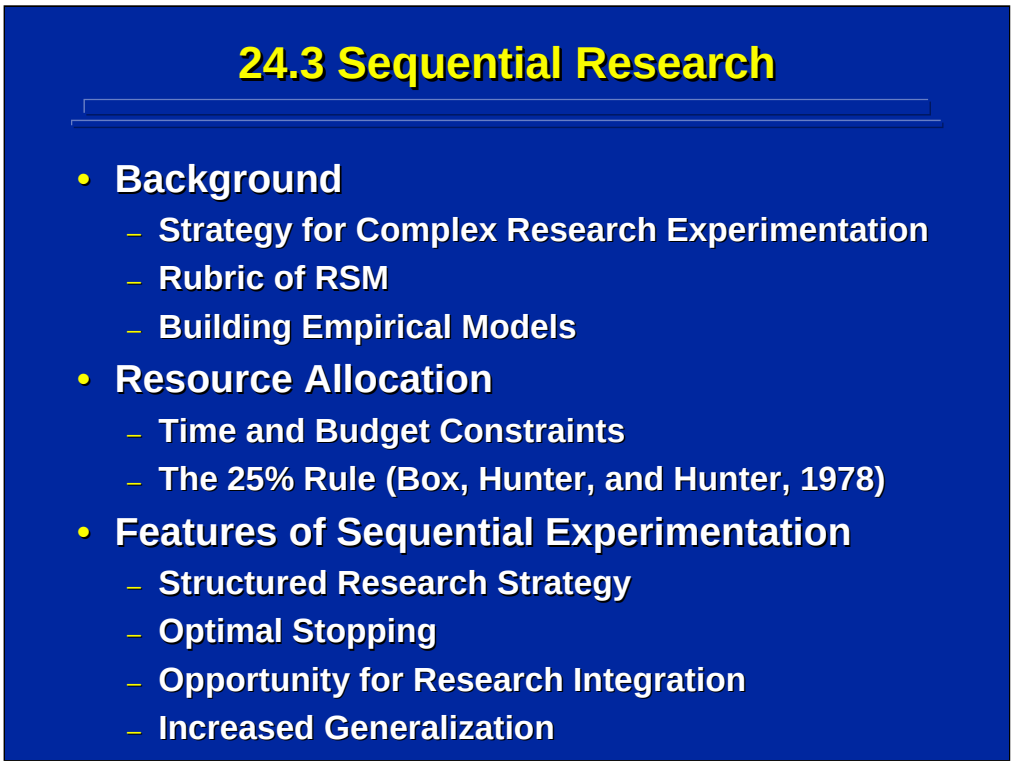



As shown on the previous slide, the method of steepest ascent is a major step in evaluating first-order designs used in RSM. Box and Draper (1987) in Chapter 6 and Myers and Montgomery (2005) in Chapter 5 provide a detailed discussion of this method.

Essentially, all of the factors of interest are used in this method to form a first-order polynomial regression that describes a hyper-plane region of the response surface. The slope of the first-order surface is rapidly ascended by changing the values of the factor levels in additional runs or in another experiment relative to the first-order partial regression weights of the previous experiment. This procedure is iterated until the partial regression weights of the first-order model are relatively constant or a significant lack of fit is achieved. This signifies approaching the maximum. Second-order designs are then used to represent the region of the maximum.




This slide summarizes the specific steps used in the method of steepest ascent as described by Myers and Montgomery (2005, p. 204). The key to this technique is the use of a series of small interrelated, first-order experiments or data runs rather than one large higher-order experiment.




Sequential experimentation, as characterized by RSM, is a useful strategy for conducting systematic research in a large data space. The series of small interrelated experiments can be combined and then used to build empirical models that predict human performance in a complex system and to conduct design tradeoffs for optimum interface design. Time and budget constraints must be allocated across this series of experiments. As a guideline, Box, Hunter, and Hunter (1978, p. 304) suggested that no more than 25% of budgeted resources be allocated to the first experiment in the series so that changes in strategy and research direction can still be made in subsequent experiments in the series.

Using a structured strategy for planning and conducting sequential experimentation is a primary key to success. Such a strategy must provide multiple opportunities for stopping and changing research direction as well as a procedure for integrating data across experiments with confidence. The result of sequential research and integration can provide a marked increase in generalization since several factors and levels of factors have been investigated across these related experiments.

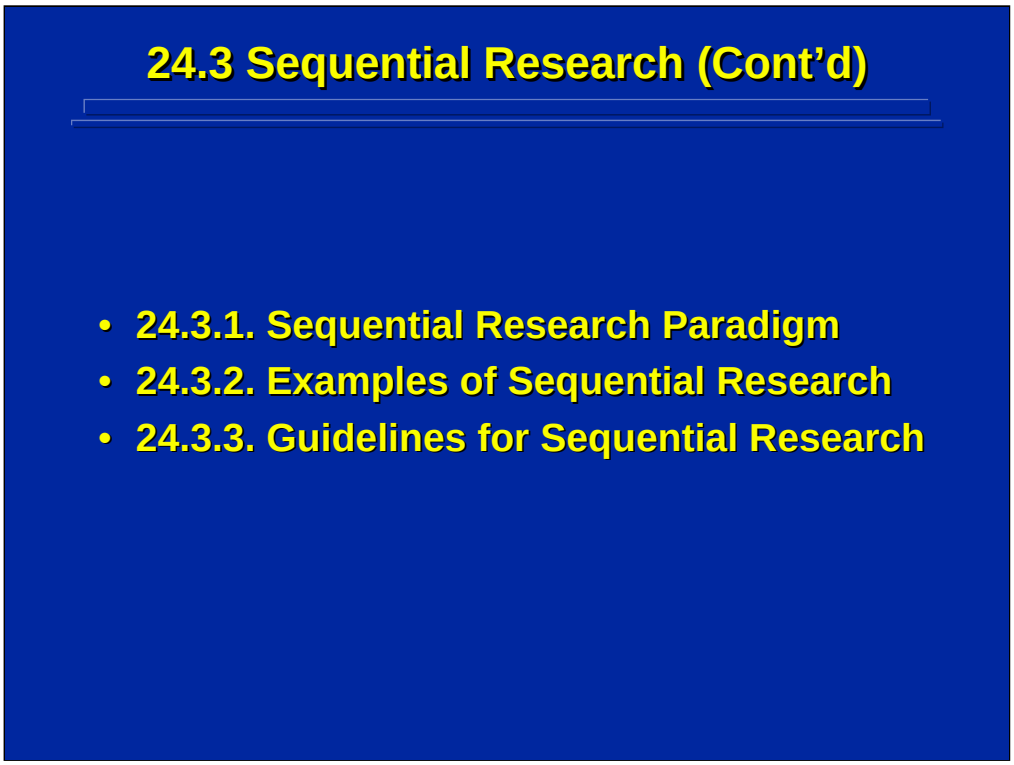
## 24.3 Sequential Research (Cont'd)

- **Sequential Research Strategy Components**
  - **Assumption: Only Main Effects and Two-Way Interactions are of Primary Interest**
  - **Selection of Data Points**
    - **Factors and Factor Levels**
    - **System Parameters**
    - **Common Data Point**
  - **Standard Experimental Procedures**
  - **Efficient Experimental Designs**
    - **$2^k$ and $2^{k-p}$ First-Order Designs**
    - **Second-Order, Central-Composite Designs**

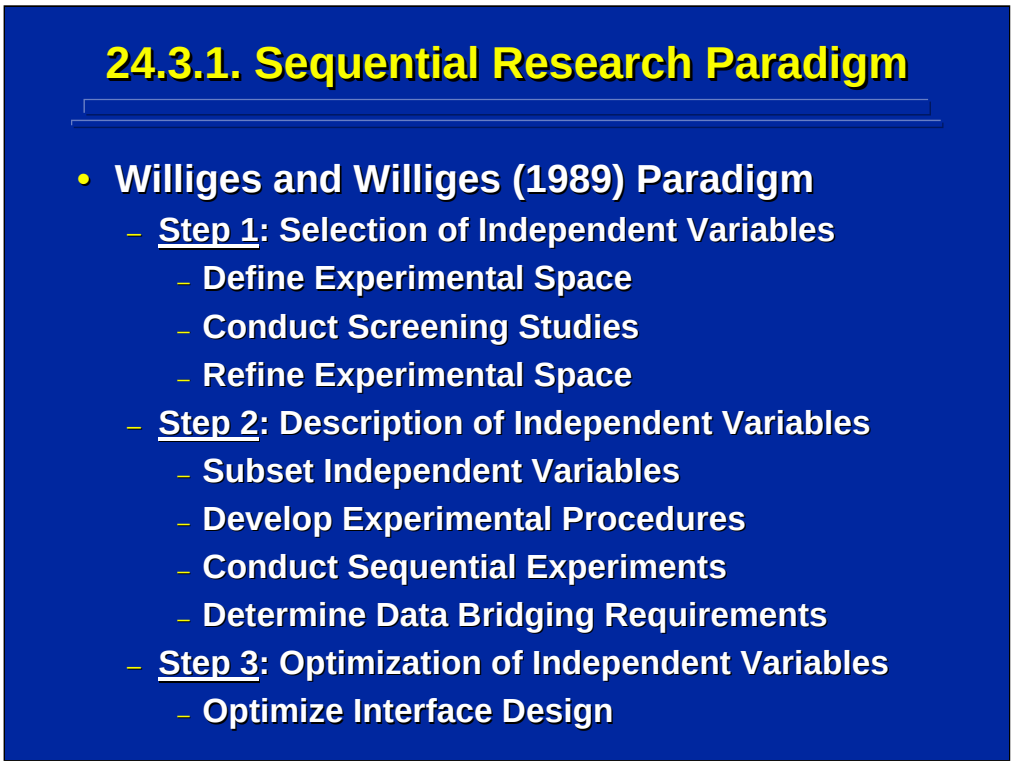
The remainder of this topic is devoted to a description of using sequential research to build empirical models of human performance in complex systems. This slide lists the four major components of this sequential research strategy. First, the experimenter assumes that only main effects and two-way interactions are of primary interest and need to be represented in the empirical model. The model is tested for lack of fit due to potential higher-order effects, but usually the model is not specified beyond a complete second-order model.

Second, a great deal of planning, screening, and pre-testing is devoted to selecting factors and levels of interest, system parameters, and a common data point observed in each study to test comparability of data for database integration. Third, standard experimental procedures for instructions, tasks, and data recording are followed to facilitate comparability across experiments.

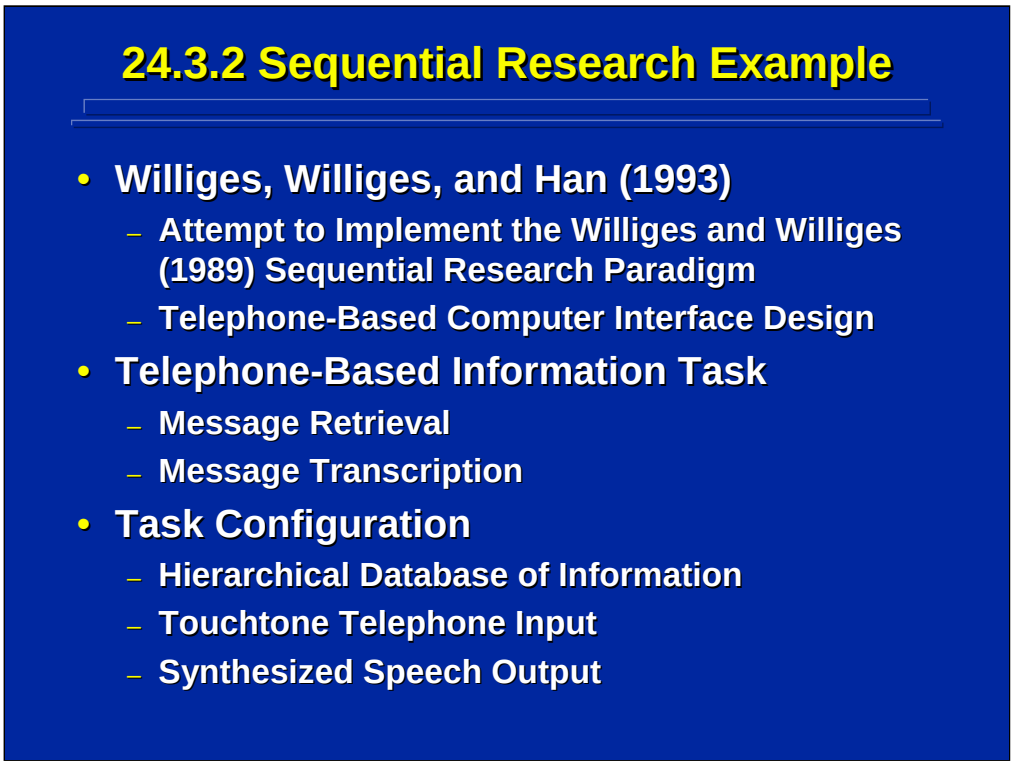
Finally, sequential research features the use of small, economical, first- and second-order experimental designs. These designs are characterized by $2^k$ and $2^{k-p}$ ANOVA designs and central-composite designs.



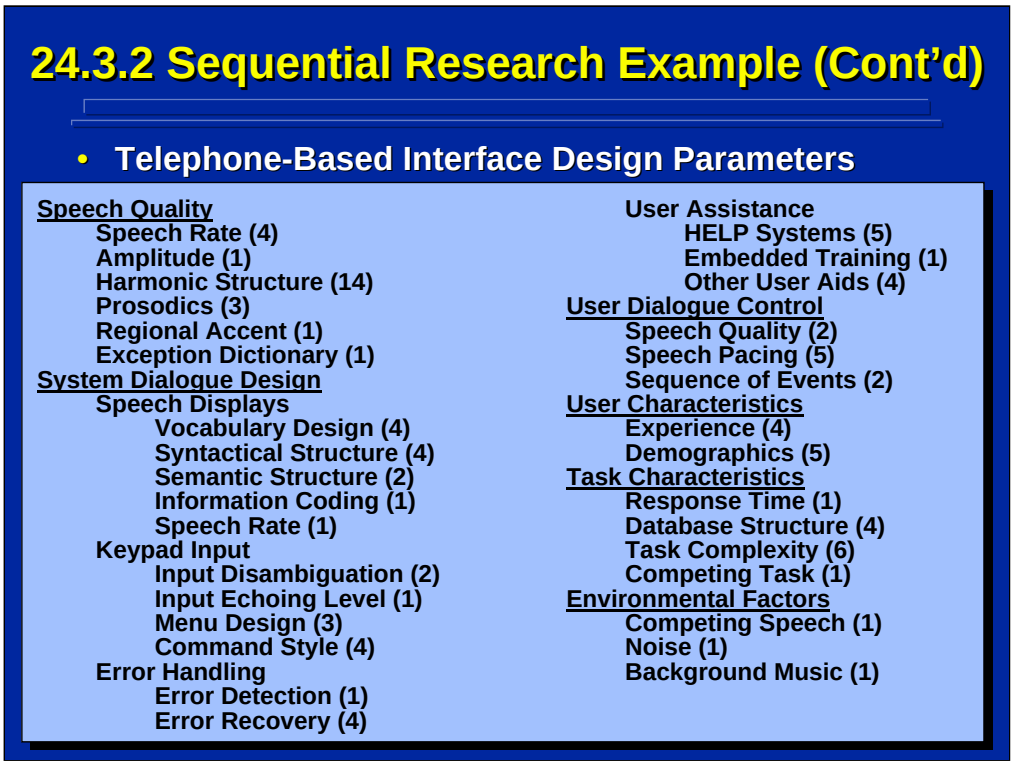

This subsection describes the use of sequential research in human factors and ergonomics. First, a general sequential research paradigm is presented followed by a detailed presentation of a human factors example using this paradigm. Finally, general guidelines for conducting sequential research are presented based on this example application.




Williges and Williges (1989) proposed a paradigm for sequential research in human factors that involves the three-step approach shown on this slide. Step 1 includes both experimental design and non-experimental design techniques to select the subset of independent variables for subsequent sequential experimentation. Step 2 involves the actual series of sequential experiments used to build the second-order empirical model that predicts human performance as a function of the independent variables of interest. Finally, Step 3 includes procedures for using the empirical model developed in Step 2 to optimize the interface design.




Williges, Williges, and Han (1993) used the Williges and Williges (1989) sequential research paradigm to investigate a telephone-based computer interface design. This interface was used by individuals to receive and transcribe information about items in a hypothetical department store. The store information was constructed in a hierarchical database that could be searched by touchtone telephone input while receiving synthesized speech output.




This slide lists 94 potential factors that can be considered in the design of the telephone-based interface. The numbers in parenthesis list the number of factors defined by the heading description. For example, there were 14 potential factors to consider in designing the harmonic structure of the speech output system.

## 24.3.2 Sequential Research Example (Cont'd)

- **Step 1: Selection of Independent Variables**
  - Initial Selection of Factors
    - Nonexperimental Design Techniques
      - Brainstorming
      - Literature Review
      - Prototyping
      - Feasibility and Relevance Analysis
      - Subjective Ratings
    - Decision Criteria
  - Conducting Screening Studies
    - Choice of Factors
    - Consider Interactions
    - Resolution IV Designs
      - 16 Factor Design (32 Versus 65,536 Observations)

Step 1 in the Williges and Williges (1989) paradigm summarized on a previous slide dealt with selecting the major independent variables that would be investigated through sequential experimentation. Both non-experimental and experimental procedures were used to select these independent variables. Merkle, Beaudet, Williges, Herlong, and Williges (1988) used the set of non-experimental procedures listed on this slide to reduce the 94 independent variables listed on the previous slide to a candidate subset of 16 factors.

A screening study was used as an experimental design procedure to select the final set of independent variables. Beaudet and Williges (1988) described a Resolution IV, $2^{16-11}$ fractional-factorial screening study consisting of 32 treatment combinations involving 16 factors. The results of their screening study further reduced the independent variables to a subset of 10 significant factors requiring further investigation. As noted at the bottom of this slide, a full $2^{16}$ factorial design would require over 64,000 treatment observations which are not feasible. Consequently, only highly economical design alternatives that resolve main effects and two-way interactions should be considered for screening experiments.

## 24.3.2 Sequential Research Example (Cont'd)

- **Ten Factors in Sequential Experimentation**

| Independent Variable | | Study | Coded Values -1.4 | -1 | 0 | +1 | +1.4 |
|---|---|---|---|---|---|---|---|
| $X_1$ | Speech Rate (SR) | 1,2,3,4 | 120 | 138 | 180 | 222 | 240 |
| $X_2$ | Input Timeout (IT) | 1,2,3,4 | 2 | 3 | 6 | 9 | 10 |
| $X_3$ | Background Music (BM) | 1,4 | 36.1 | 40.5 | 50.8 | 61.1 | 65.4 |
| $X_4$ | Age of User (AU) | 1,4 | 15 | 22 | 38 | 54 | 60 |
| $X_5$ | Menu Structure (MS) | 2,4 | | $2^6$ | | $8^2$ | |
| $X_6$ | Feedback (FB) | 2,4 | | No | | Yes | |
| $X_7$ | User Guide (UG) | 2,4 | | No | | Yes | |
| $X_8$ | No. of Messages (NM) | 3,4 | | 1 | | 2 | |
| $X_9$ | Gender of User (GU) | 3,4 | | Female | | Male | |
| $X_{10}$ | Type of Voice (TV) | 3,4 | | Betty | | Paul | |

Reprinted from Williges, Williges, and Han (1993) by Permission

This slide summarizes the final subset of the 10 independent variables investigated through sequential research procedures. The series of experiments used in this example involved four separate data collection studies. Note that Speech Rate and Input Timeout factors (i.e., $X_1$ and $X_2$, respectively) are considered parameters because they were manipulated in all four sequential research studies.

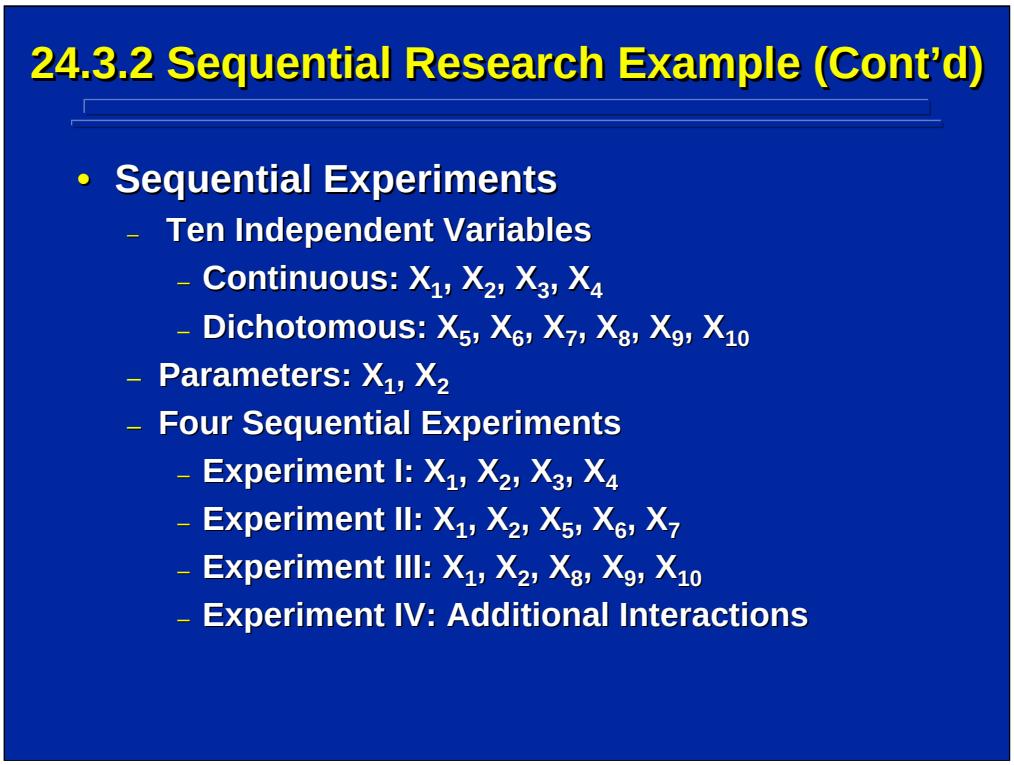
The first four factors were manipulated at five levels across the sequential experiments; whereas, the other six factors were only manipulated at two levels. Both coded values and real-world levels of the 10 factors are shown on the slide.

## 24.3.2 Sequential Research Example (Cont'd)

- **Step 2: Description of Independent Variables**
  - Series of Small Experiments
    - $2^k$ Factorial Designs
    - $2^{k-p}$ Resolution V Fractional-Factorial Designs
    - Central-Composite Designs
  - Empirical Model Building
    - Polynomial Regression
  - Data Bridging Across Experiments
  - Integrated Empirical Model

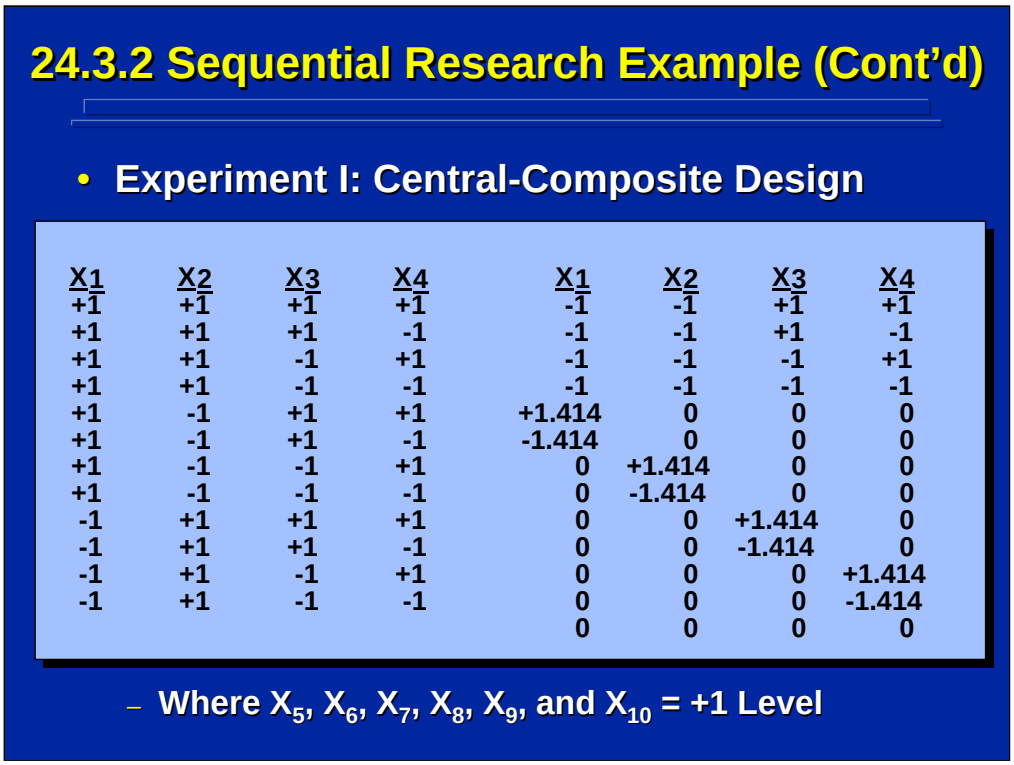
Step 2 of the Williges and Williges (1989) paradigm as shown on a previous slide deals with describing inter-relationships among the 10 factors selected in Step 1. Step 2 is the actual set of sequential research studies. This sequence is a series of small experiments that provide separate, meaningful results and can generate empirical models based on polynomial regression that, at least, represent main effects and two-way interactions of the factors manipulated in each experiment. Generally, $2^k$ factorial designs, $2^{k-p}$ fractional factorial designs of Resolution V, and central-composite designs are the main candidates for this series of small studies. These separate experiments investigate meaningful groupings of independent variables and stand alone as experiments that can generate empirical models of the subset of factors investigated.

Data bridging is a key component of the Step 2 sequential research process. Only additional data points needed to resolve all the two-way interactions among the factors of interest are collected in data bridging. Consequently, data bridging usually consists of a small collection of additional observations rather than a completely balanced experimental design. These additional data points can be combined with the data collected in the preceding set of small experiments to form the database for building an integrated empirical model that can include all the main factors of interest.




This slide summarizes the series of sequential studies in the example problem that were conducted in Step 2 of the Williges and Williges (1989) paradigm. The 10 factors in the example problem consisted of four continuous factors and six dichotomous factors. Two of these ten factors, $X_1$ and $X_2$, were parameters investigated in each experiment in the sequence.

A sequence of four separate data collections was conducted. As shown on the slide, Experiment I involved factors $X_1$, $X_2$, $X_3$, and $X_4$; Experiment II investigated factors $X_1$ and $X_2$ with factors $X_5$, $X_6$, and $X_7$; Experiment III evaluated factors $X_1$ and $X_2$ with factors $X_8$, $X_9$, and $X_{10}$; and Experiment IV manipulated all ten factors in the data bridging.




Experiment I in the example problem was a four factor, central-composite design (CCD). The coded values of the 25 treatment combinations of the CCD are shown on this slide. An $\alpha$ value of 1.414 was used to provide an orthogonal CCD as described in Topic 23. The remaining six factors of interest that were not manipulated in this experiment were held constant at the +1 coded level as noted at the bottom of this slide.

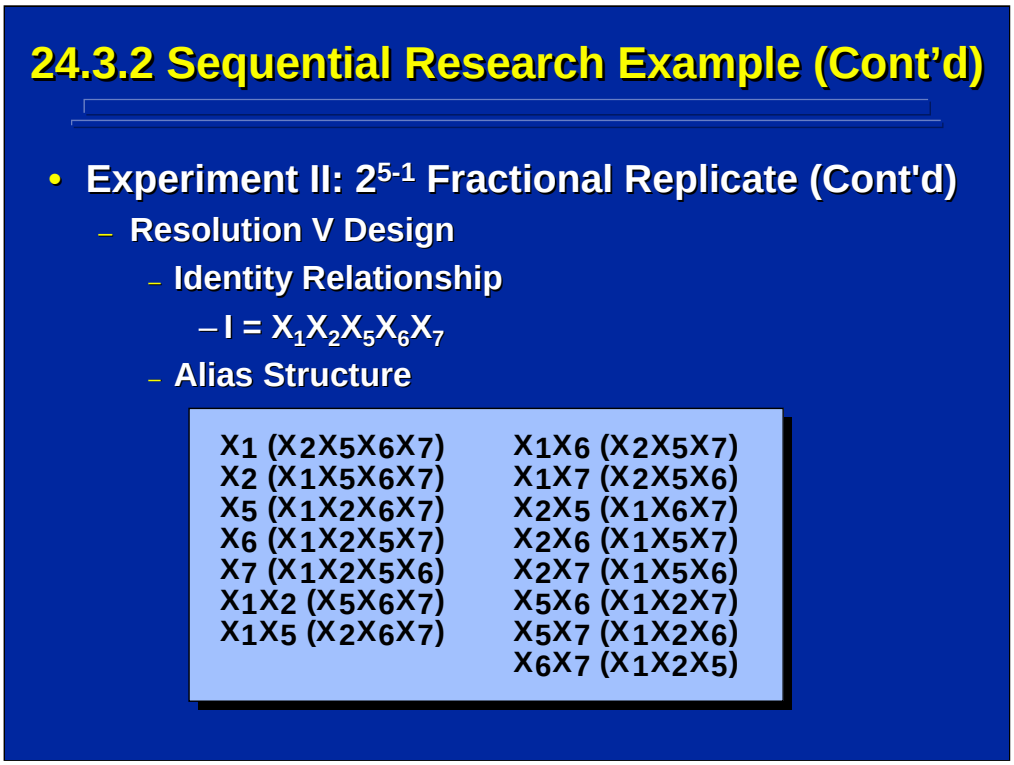
## 24.3.2 Sequential Research Example (Cont'd)

- **Experiment II: $2^{5-1}$ Factional Replicate**

| $X_1$ | $X_2$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| -1.414 | -1.414 | -1 | -1 | +1 |
| -1.414 | -1.414 | -1 | +1 | -1 |
| -1.414 | -1.414 | +1 | -1 | -1 |
| -1.414 | -1.414 | +1 | +1 | +1 |
| -1.414 | +1.414 | -1 | -1 | -1 |
| -1.414 | +1.414 | -1 | +1 | +1 |
| -1.414 | +1.414 | +1 | -1 | +1 |
| -1.414 | +1.414 | +1 | +1 | -1 |
| +1.414 | -1.414 | -1 | -1 | -1 |
| +1.414 | -1.414 | -1 | +1 | +1 |
| +1.414 | -1.414 | +1 | -1 | +1 |
| +1.414 | -1.414 | +1 | +1 | -1 |
| +1.414 | +1.414 | -1 | -1 | +1 |
| +1.414 | +1.414 | -1 | +1 | -1 |
| +1.414 | +1.414 | +1 | -1 | -1 |
| +1.414 | +1.414 | +1 | +1 | +1 |

  - **Where $X_3$, $X_4$, $X_8$, $X_9$, and $X_{10}$ = +1 Level**

Experiment II in the example problem was a Resolution V, $2^{5-1}$ fractional-factorial design. The coded values of the five factors in this design are shown on this slide. Note that the ±1.414 levels of $X_1$ and $X_2$ were chosen even though any two of the five levels of these two factors could have been used. The remaining five factors of interest that were not manipulated in this experiment were held constant at the +1 coded level as noted at the bottom of this slide.




The design used in Experiment II is a one-half replicate of a $2^5$ factorial design. The identity relationship used in the one-half replicate is shown on this slide and uses the five-way interaction as described in Topic 18. This identity relationship was chosen to form a Resolution V design that keeps the main effects and two-way interactions of the five factors orthogonal from each other so that the resulting empirical model based on the results of this experiment could represent unconfounded partial regression weights of the main effects and two-way interactions. The resulting alias structure for this design is shown at the bottom of this slide.

## 24.3.2 Sequential Research Example (Cont'd)

- **Experiment III: $2^{5-1}$ Fractional Replicate**

| $X_1$ | $X_2$ | $X_8$ | $X_9$ | $X_{10}$ |
|---|---|---|---|---|
| -1 | -1 | -1 | -1 | +1 |
| -1 | -1 | -1 | +1 | -1 |
| -1 | -1 | +1 | -1 | -1 |
| -1 | -1 | +1 | +1 | +1 |
| -1 | +1 | -1 | -1 | -1 |
| -1 | +1 | -1 | +1 | +1 |
| -1 | +1 | +1 | -1 | +1 |
| -1 | +1 | +1 | +1 | -1 |
| +1 | -1 | -1 | -1 | -1 |
| +1 | -1 | -1 | +1 | +1 |
| +1 | -1 | +1 | -1 | +1 |
| +1 | -1 | +1 | +1 | -1 |
| +1 | +1 | -1 | -1 | +1 |
| +1 | +1 | -1 | +1 | -1 |
| +1 | +1 | +1 | -1 | -1 |
| +1 | +1 | +1 | +1 | +1 |

  - **Where $X_3$, $X_4$, $X_5$, $X_6$, and $X_7$ = +1 Level**

Experiment III in the example problem also used a Resolution V, $2^{5-1}$ fractional-factorial design. The same alias structure shown on the previous slide for Experiment II was also used in this experiment.

The coded values of the five factors in this design are shown on this slide. Note that the ±1 levels of $X_1$ and $X_2$ were chosen even though any two of the five levels of these two factors could have been used. The remaining five factors of interest that were not manipulated in this experiment were held constant at the +1 coded level as noted at the bottom of this slide.

## 24.3.2 Sequential Research Example (Cont'd)

- **Experiment IV: Data Bridging**
  - **Data Bridging Procedure (Han, Williges, and Williges, 1997)**
    - **Data Integration**
    - **Data Point Selection**
    - **Check for Multicollinearity**
  - **16 Unresolved Two-Way Interactions**

    **$X_3X_5$, $X_3X_6$, $X_3X_7$, $X_3X_8$, $X_3X_{10}$, $X_4X_5$, $X_4X_6$, $X_4X_7$, $X_4X_8$, $X_4X_{10}$, $X_5X_8$, $X_5X_{10}$, $X_6X_8$, $X_6X_{10}$, $X_7X_8$, and $X_7X_{10}$**
  - **Data Collection Requirement**
    - **One Data Point for Each Interaction**
    - **$2^8$ Alternatives for Each Data Point**

The fourth experiment in the example problem is not really an experiment in terms of using an experimental design. Rather, it is just a collection of additional data points needed to resolve all two-way interactions across the entire set of 10 factors. Han, Williges, and Williges (1997, pp 572-575) describe the details of the procedures for determining treatment combinations needed in data bridging that involve the three considerations of data integration, data point selection, and a check for multicollinearity when the regressors are truly not independent.

Data integration requires a determination of effects that still need to be evaluated when combining the data across previous experiments in the research sequence. As listed on this slide, there are 16 two-way interactions in the example problem that were not investigated in the previous three experiments in the research sequence. For each of these interactions, the experimenter must collect one more data point in order to evaluate the interaction. Since there are 8 other factors each at two levels, there are $2^8$ alternatives for choosing each additional data point.

## 24.3.2 Sequential Research Example (Cont'd)

- **Experiment IV: Data Bridging (Cont'd)**
  - **Mathematical Selection Criterion**
    - **Maximize Determinant of X'X Matrix**
    - **Reduce Degree of Multicollinearity**
  - **Reduced to Six Required Data Points**

| Data Point | Factor Levels of Six Additional Treatment Conditions | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ |
| 1 | -1 | +1 | +1 | -1 | -1 | +1 | +1 | +1 | +1 | -1 |
| 2 | +1 | -1 | +1 | +1 | -1 | +1 | +1 | +1 | +1 | -1 |
| 3 | +1 | +1 | +1 | +1 | +1 | +1 | +1 | +1 | -1 | -1 |
| 4 | -1 | -1 | +1 | -1 | -1 | +1 | -1 | +1 | +1 | -1 |
| 5 | +1 | -1 | +1 | -1 | +1 | +1 | +1 | +1 | +1 | -1 |
| 6 | -1 | +1 | +1 | +1 | -1 | +1 | -1 | +1 | +1 | -1 |

Reprinted from Williges, Williges, and Han (1993) by Permission

Han, Williges, and Williges (1997) described a mathematical selection and evaluation procedure based on the mathematical criterion of maximizing the determinate of the X'X matrix and checking for multicollinearity as described by Myers (1990) in Chapter 8. By following this procedure, the data bridging was reduced to six additional data points. The coded values of the 10 factors for these six additional data points are listed at the bottom of this slide.

## 24.3.2 Sequential Research Example (Cont'd)

- **Integrated Empirical Models**

**TST = 52.81 + 17.27(IT) + 8.27(MS) + 4.49(ITxMS) - 4.47(UG)**
**+ 2.83(MSxUG) + 2.77(BM) + 2.15(AU) + 1.66(ITxBM)**
**- 1.55(ITxUG) - 1.20(SR)**

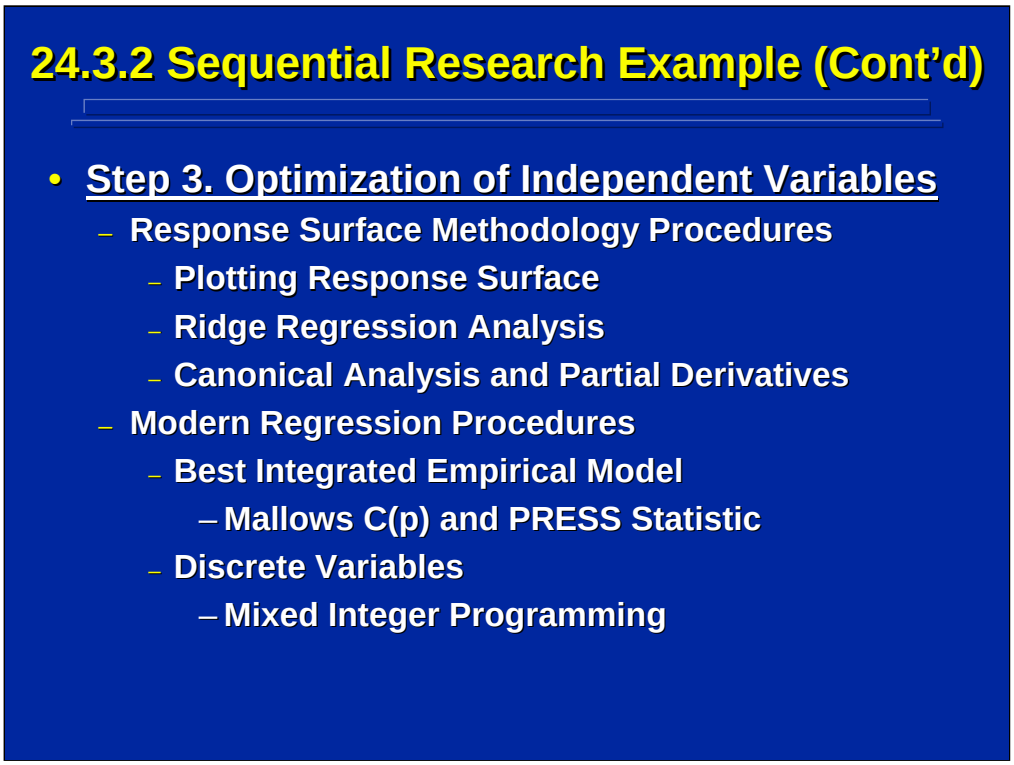
**UAK = 1.02 - 0.34(MS) + 0.28(AU) + 0.23(MSxUG) + 0.21(AUxUG)**
**+ 0.13(BM) + 0.12(NM) - 0.11(UG) - 0.07(ITxUG)**

**TA = 3.16 - 0.40(BM) - 0.21(BM $^{2}$) - 0.15(SR) - 0.15(AU )**
**- 0.02(ITxGU)**

**TST = Total Search Time**
**UAK = User Added Key Presses**
**TA = Transcription Accuracy**

**SR = Speech Rate**
**IT = Input Timeout**
**BM = Background Music**
**AU = Age of User**
**MS = Menu Structure**
**UG = User Guide**
**NM = Number of Messages**
**GU = Gender of User**

The final aspect of Step 2 of the Williges and Williges (1989) paradigm that was shown on a previous slide is to build integrated empirical models based on all the data collected in the sequential studies. This slide summarizes three integrated models of all 10 factors in the example problem that separately represent the three major dependent variables, Total Search Time (TST), User Added Key Presses (UAK), and Transcription Accuracy (TA). Only significant ($p < 0.05$) first-order and second-order effects are included in these empirical models.




The final step, Step 3, of the Williges and Williges (1989) paradigm deals with optimizing performance across the sequential experiments to determine the best combination of independent variables to define the interface. Both Response Surface Methodology (RSM) and modern regression procedures can be used to aid in optimization.

As shown on the top of this slide, various RSM procedures such as plots of response surfaces, ridge regression, and canonical analysis can be used to provide a good description of the region of optimality. In addition, partial derivatives of the integrated empirical model can be considered as a means of defining the optimal level of factor combinations.

Modern regression procedures using Mallows C(p) and the PRESS statistic can be used to select the best integrated model when covariance exists among the partial regression weights. Often in human factors research, some of the factors are a combination of continuous and discrete variables, and mixed integer programming may need to be considered for optimization purposes.

## 24.3.2 Sequential Research Example (Cont'd)

- **Optimization Procedures**
  - Nonlinear Search on Continuous Variables
  - All Combinations of Dichotomous Variables
  - Summary of Optimum Interface Configuration

| Dependent Variables | Ten Independent Variables | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | SR | IT | BM | AU | MS | UG | NM | FB | GU | TV |
| TST | 1.4 | -1.4 | -1.4 | -1 | -1 | 1 | ** | ** | ** | ** |
| UAK | ** | -1.4 | -1.4 | -1 | 1 | 1 | -1 | ** | ** | ** |
| TA | -1.4 | -1.4 | -0.9 | -1 | ** | ** | ** | ** | ** | ** |

** Any Value Within The Design Region

TST = Total Search Time
UAK = User Added Key Presses
TA = Transcription Accuracy

SR = Speech Rate
IT = Input Timeout
BM = Background Music
AU = Age of User
MS = Menu Structure
UG = User Guide
NM = Number of Messages
FB = Feedback
GU = Gender of User
TV = Type of Voice

Williges, Williges, and Han (1993) used a nonlinear search of the continuous variables and all combinations of the dichotomous variables. This yielded the best approach in determining the optimal performance in the example problem. This slide shows the results of these procedures to optimize the three integrated empirical models shown on a previous slide. The optimal level for the 10 factors on each of the three major dependent variables is shown on this slide. Note that any level can be used with some of the variables.

These treatment combinations can be used to specify the optimal telephone-based interface configuration of the 10 independent variables investigated. Since each of the three dependent variables results in a slightly different optimum configuration, the experimenter must trade off these dependent variables to optimize the telephone-based interface.

## 24.3.2 Sequential Research Example (Cont'd)

- Quantitative Design Guidelines
  - Optimum Interface Configuration
  - Interface Design Tradeoffs
    - Functional Relationships
    - Alternative Configurations
  - Quantitative Design Impact
    - Functional Relationship
    - Magnitude of Effect
    - Direction of Influence
- Qualitative Design Guidelines
  - Verbal Descriptions
  - Interface Design Rules

Empirical models can be used in human factors research to provide both quantitative and qualitative interface design guidelines. Since empirical models are quantitative prediction equations, most applications are quantitative in nature. As noted on this slide, the optimum interface configuration can be defined and design tradeoffs can be evaluated. In addition, design impacts of changing various factor levels can be evaluated based on the empirical model itself. Considerations such as predicting performance based on the functional relationship, determining the relative magnitude of orthogonal partial regression weights, and understanding the direction of influence of the various factors can aid in the human factors interface design process.

Qualitative interface design guidelines based on research data are also of major interest to the human factors researcher. These verbal descriptions need to be developed based on the relationships specified in the empirical models resulting from sequential research, and they can provide interface design rules. These rules are validated through accepted use and follow-on research.

## 24.3.2 Sequential Research Example (Cont'd)

**Qualitative Design Guidelines**

**(Reprinted from Williges, Williges, & Han, 1992, by Permission)**

1. **Auditory menus should be short with few menu alternatives.**
2. **Methods to reduce the presentation time of auditory menus should be provided. This may include allowing the user to Interrupt speech displays or menus or to stack input commands.**
3. **The time between the presentation of auditory menu alternatives should be minimal and should not exceed 2 seconds.**
4. **For auditory databases that do not change frequently, a flat database structure should be used for immediate access. Each item should be assigned an access code, and a user guide to these codes should be distributed.**
5. **After completion of each auditory search, the system should return the user to the top of the menu hierarchy before another search can be initiated.**

As an example of generating qualitative design guidelines, Williges, Williges, and Han (1992) provide ten interface design guidelines for the telephone-based interface that are based on the empirical models generated in the example problem through sequential experimentation. The first five guidelines are presented on this slide. (Reprinted with permission from *Human Factors*. Copyright 1992 by the Human Factors and Ergonomics Society. All rights reserved.)

## 24.3.2 Sequential Research Example (Cont'd)

**Qualitative Design Guidelines**

**(Reprinted from Williges, Williges, & Han, 1992, by Permission)**

6. **Throughout the search process a simple method should be provided to allow the user to return to the top of an auditory menu structure.**
7. **An auditory information system should minimize the effects of noise or background music on the intelligibility of the auditory displays.**
8. **If an auditory information system will be used by older adults, the auditory displays should be designed to maximize speech intelligibility to offset the reduction in speech intelligibility with age.**
9. **If an auditory information system will be used by older adults, waiting periods for menu selections should be increased to account for the increase in choice reaction time.**
10. **The rate of rule-based synthesized speech should be set at a mid-level such as 180 words per minute.**

This slide summarizes the second five qualitative interface design guidelines specified by Williges, Williges, and Han (1992). (Reprinted with permission from *Human Factors*. Copyright 1992 by the Human Factors and Ergonomics Society. All rights reserved.)

No rules are available for specifying these verbal descriptions and interface design rules. They ultimately depend on the experience and understanding of the experimenter to make valid interpretations of the empirical results and to specify qualitative design guidelines that provide useful interface design rules. Nonetheless, these qualitative design guidelines are a necessary component of human factors and ergonomics use of sequential experimentation and should always be part of the sequential research process.

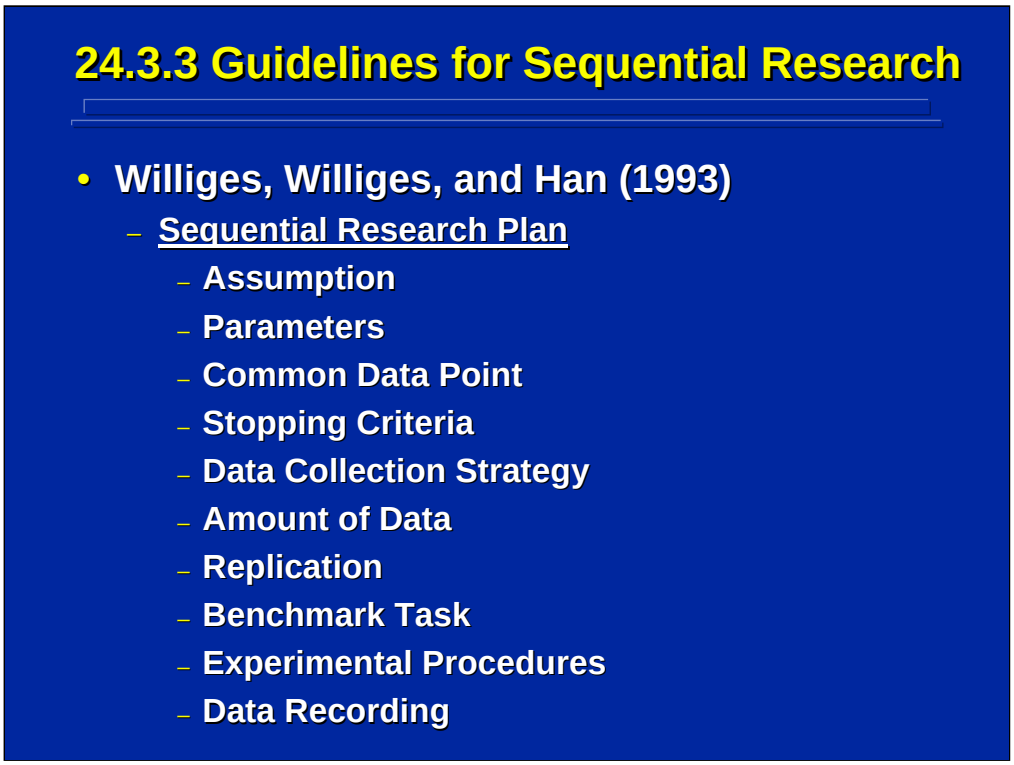
## 24.3.2 Sequential Research Example (Cont'd)

- Evaluation of Sequential Experiments
  - Economy of Sequential Research
    - Single Experiment
      - $5^4 \times 2^6 =$ 40,000 Unique Data Points
    - Four Sequential Experiments
      - $25 + 2^{5-1} + 2^{5-1} + 6 =$ 63 Unique Data Points
  - Staging of Data Collection
    - Feasible Size of Experiments
    - Multiple Stopping Points
  - Combined Data Set
    - Integrated Empirical Model
    - Interface Design Tradeoffs
    - Increased Generality

Williges, Williges, and Han (1993) evaluated their sequential research example in terms of economy of data collection, staging of data collection, and use of the combined data set. Sequential research was extremely economical in terms of investigating the various levels of the 10 factors. A total of 40,000 unique data points would be required in one large factorial experiment as compared to only 63 unique data points in the four sequential experiments.

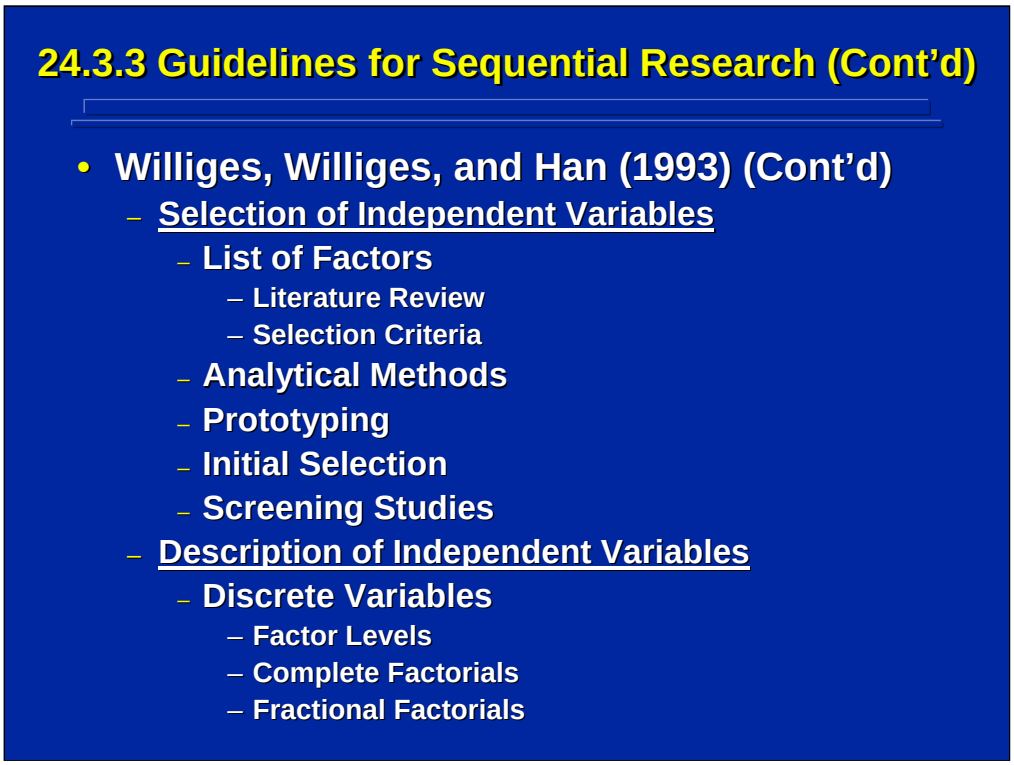
Obviously, conducting one large experiment is not feasible and staging the data collection through sequential research makes the investigation of this example complex research problem feasible. Research staging results in small experiments that are both feasible and meaningful. This staging also allows several opportunities to terminate data collection or change directions, if necessary.

Combining data across the sequential experiments allowed the development of an integrated empirical model that can be used to make design tradeoffs to result in an optimum interface design configuration. The overall generality of these results can be enhanced through the use of both quantitative and qualitative design guidelines.

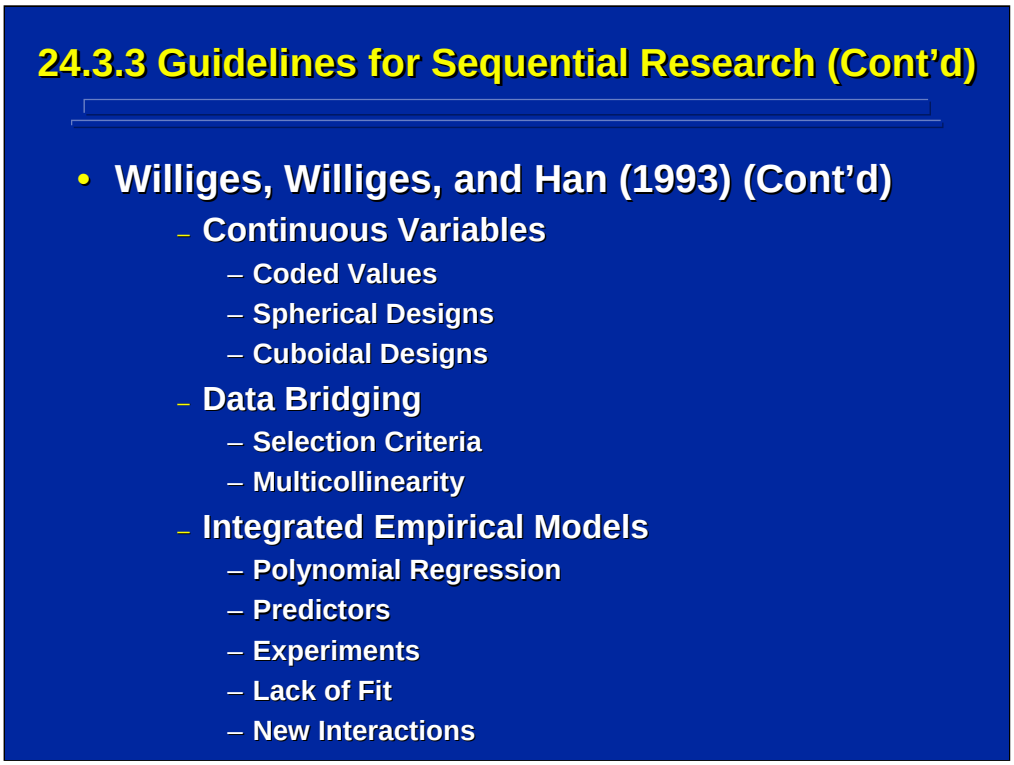



Based on their experience with sequential research as described in the previous example problem dealing with the telephone-based interface, Williges, Williges, and Han (1993, pp. 23-27) provided several design guidelines for researchers to consider when using sequential experimentation. As shown on this slide, they expanded the original Williges and Williges (1989) sequential research paradigm by adding a planning component at the beginning of sequential research and planning continues interactively throughout each additional stage of the process.

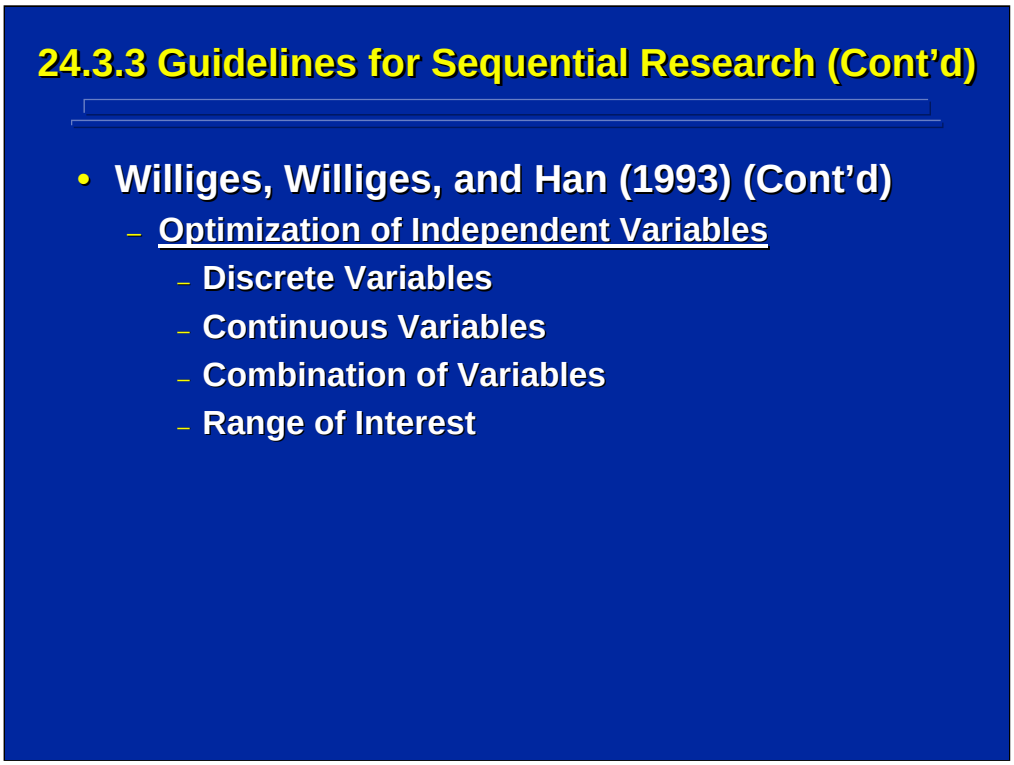
Williges, Williges, and Han (1993) presented sequential research planning guidelines that cover the ten topics shown on this slide. These ten guidelines can be used as a checklist for experimenters who are planning a series of sequential experiments.



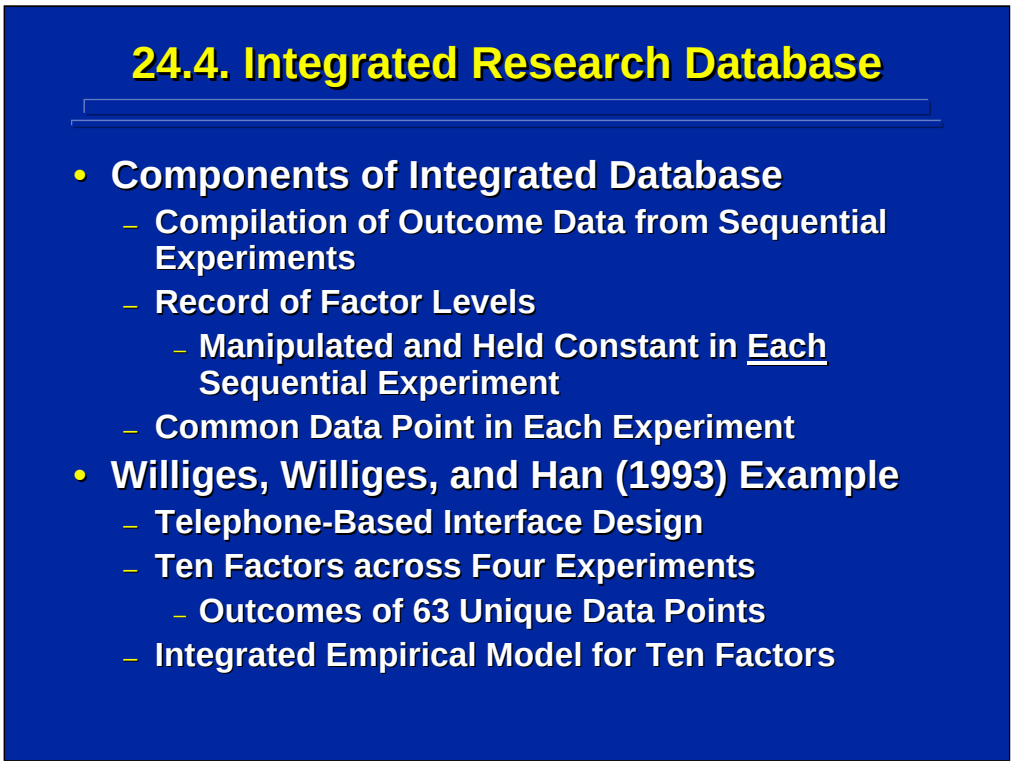

The remaining guidelines presented by Williges, Williges, and Han (1993) relate to selecting, describing, and optimizing independent variables according to the Williges and Williges (1989) paradigm. The top of this slide summarizes their guideline topics for selecting independent variables to include in sequential research. The main guideline considerations deal with the list of factors, analytical methods, prototyping, initial screening, and screening studies involved in selecting independent variables. The bottom of this slide lists guideline considerations in describing discrete variables used in sequential experimentation.




This slide continues with the Williges, Williges, and Han (1993) list of guidelines for describing independent variables in sequential research. Besides presenting guidelines for discrete variables as noted on the previous slide, this slide lists their major guideline topics for continuous variables, data bridging, and integrated empirical models that need to be considered in this stage of the sequential research process.




This slide lists the Williges, Williges, and Han (1993) guideline topics to consider in the final stage of the sequential research process. They provide guidelines for dealing with discrete variables, continuous variables, a combination of both variables, and the range of interest across variables during the optimization stage of the Williges and Williges (1989) paradigm.




One of the major outputs of sequential experimentation is the resulting integrated database that combines the results of each experiment in the sequential series. The top of this slide lists two key components of an integrated database. First, the experimenter must record the levels of each factor manipulated in each experiment and must record the level of each factor held constant in each experiment. Consequently, the resulting integrated database lists the dependent variables outcome and the level of every factor for each observation. Second, a common data point should also be observed in each experiment to determine if the results of separate experiments are compatible and can be combined into the integrated database.

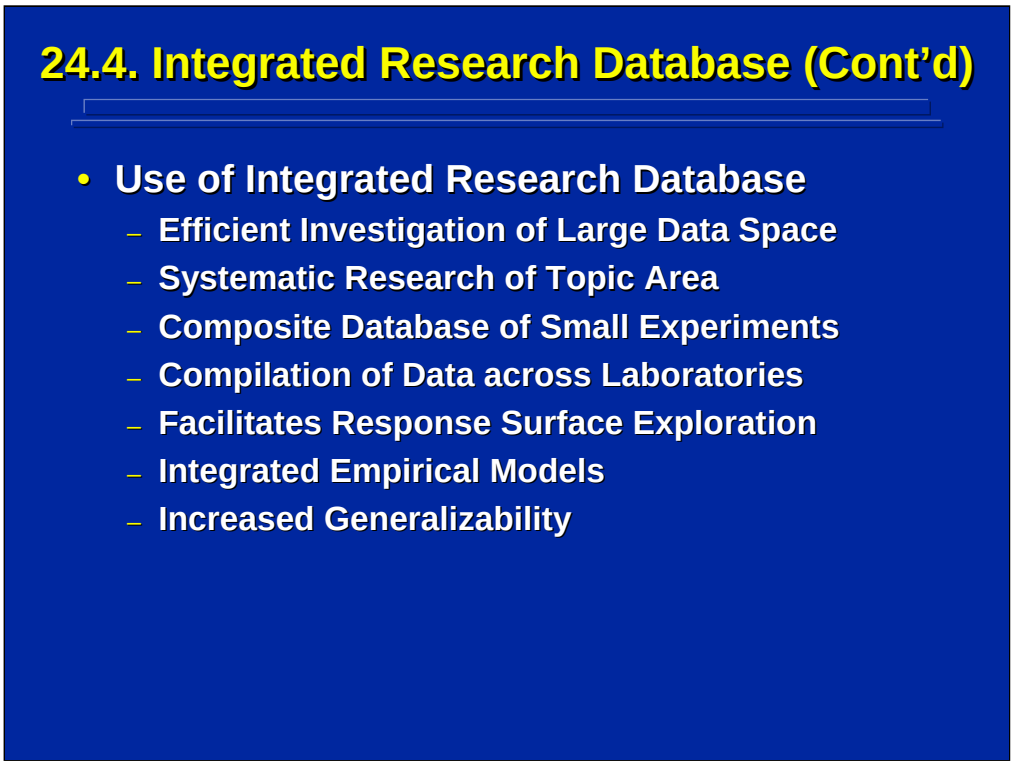
The previous example of sequential research conducted by Williges, Williges, and Han (1993) on the telephone-based interface design problem resulted in an integrated database of outcomes representing 63 unique data points collected across a series of four small experiments. This integrated database was used to generate empirical models representing the first-order and second-order effects of ten factors.

## 24.4. Integrated Research Database (Cont'd)

- Snow and Williges (1998) Example
  - Perceived Presence in Virtual Environments
  - Eleven Factors across Three Experiments
    - Outcomes of 52 Unique Data Points
      - Factors 1 to 3: 3x2x3 Factorial Design
      - Factors 4 to 8: $2^{5-1}$ Fractional-Factorial Design (Resolution V)
      - Factors 9 to 11: 3x2x3 Factorial Design
  - Common Data Point
    - Third Experiment Incompatible
  - Integrated Empirical Model of Eight Factors
    - Standardized Regression for Relative Comparison

Snow and Williges (1998) provided another human factors example of generating an integrated database resulting from sequential experimentation on eleven factors affecting operator's perceived presence in a virtual environment. As shown on the center of this slide, outcomes on a total of 52 unique data points were collected across three separate, sequential experiments.

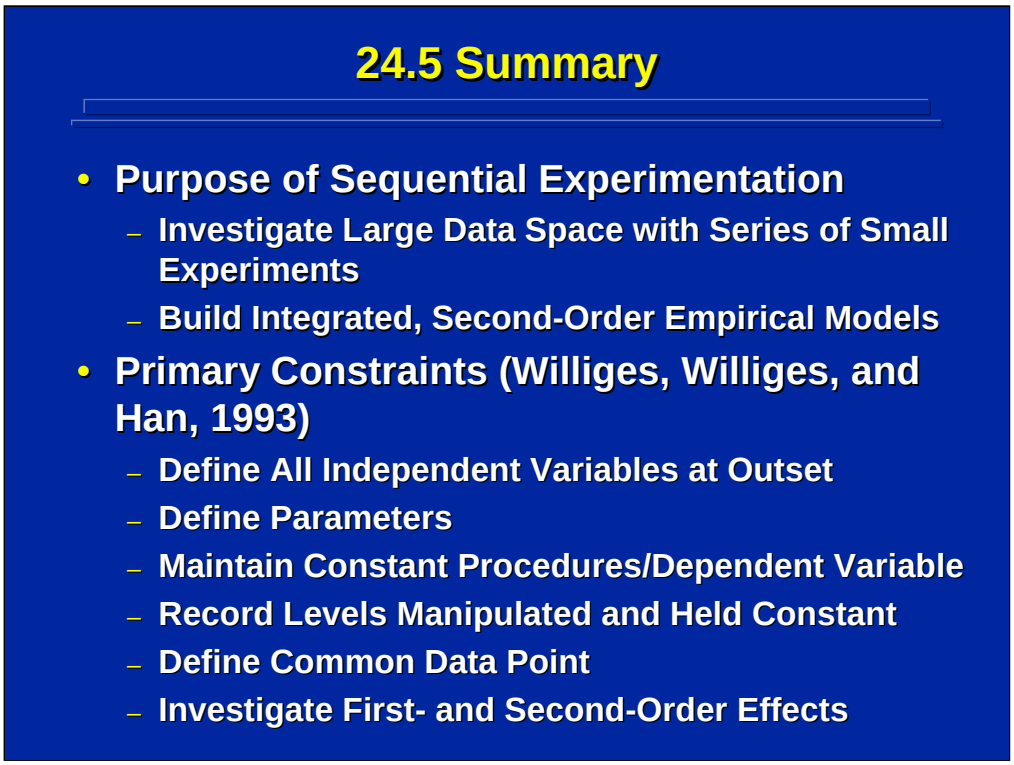
Subsequent investigation of the common data point in the three experiments showed that the results of the third experiment were significantly lower than the first two experiments. Therefore, the results of the third experiment were deemed incompatible with the other two experiments and were not included in the integrated database. Consequently, the integrated empirical model generated by Snow and Williges (1998) was restricted to the eight factors investigated in the first two sequential experiments. They then developed a standardized polynomial regression equation so that relative contributions of these factors could be evaluated.




This slide lists the various characteristics and potential uses of integrated research databases. The results are collected through efficient investigations of large data spaces and are based on well-planned, systematic studies.

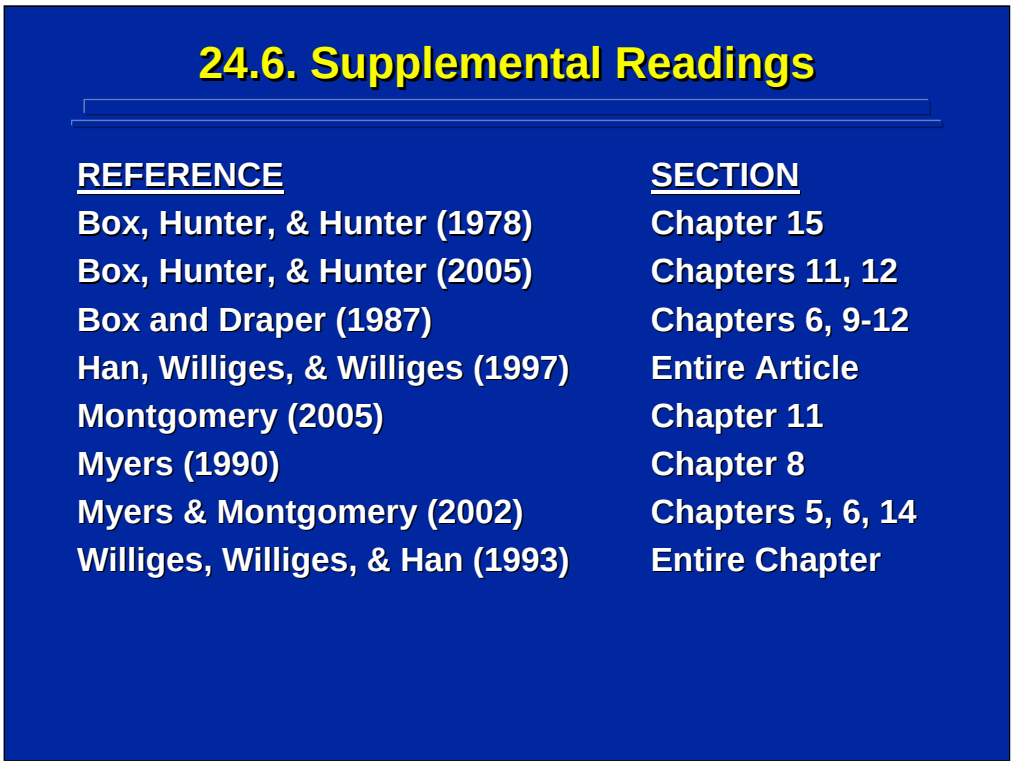
The database is a composite of small sequential experiments usually conducted in the same laboratory by the same investigators. Alternatively, the data could be compiled across laboratories provided appropriate checks are made for comparability of results and common tasks are used.

The integrated database facilitates exploration of the entire response surface by including more factors simultaneously than possible in any individual experiment in the sequential series. The resulting integrated models are more generalizable because they include more factors and a wider range of levels across factors.

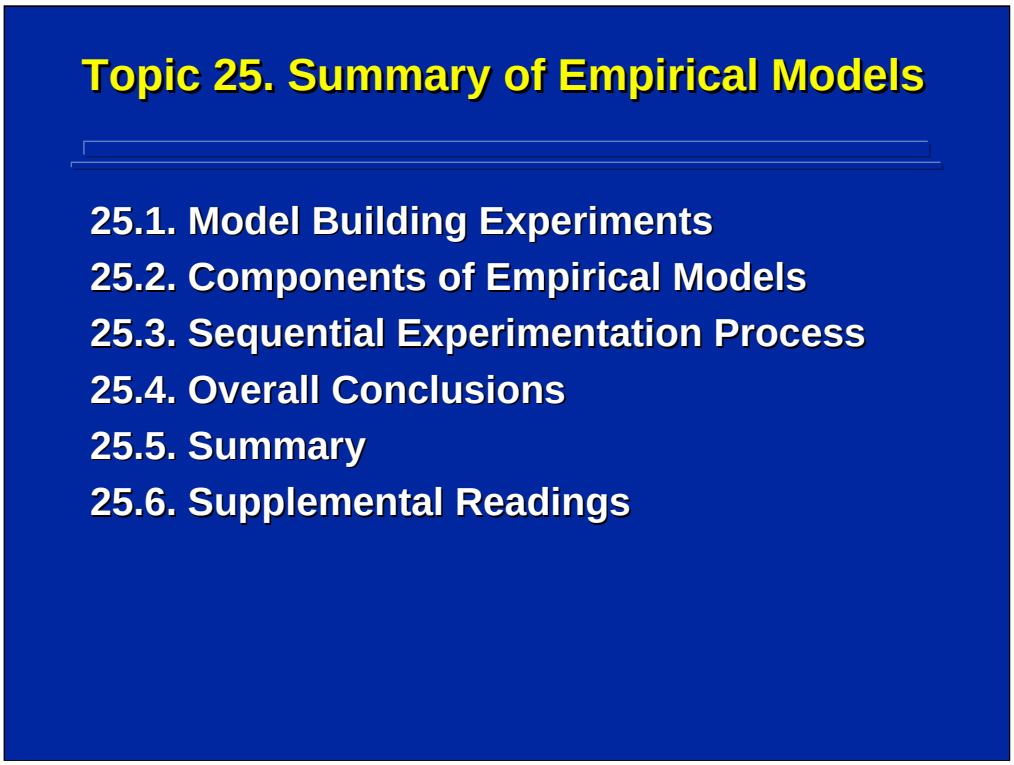



Topic 24 deals with techniques for conducting sequential experimentation to investigate large experimental spaces through a series of small experiments. Many procedures used in response surface methodology are useful in sequential experimentation. The primary purpose of sequential research is to build integrated database that can be used to generate second-order empirical models.

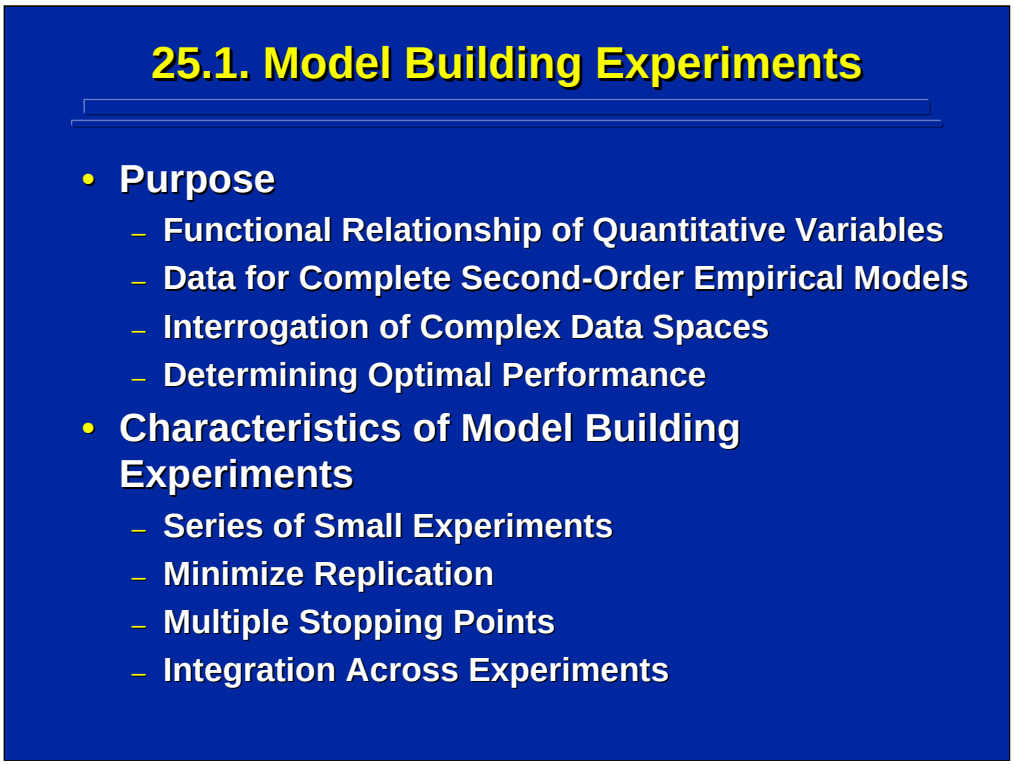
Many complex system problems in human factors and ergonomic research are amenable to sequential research. Williges, Williges, and Han (1993) recommended the six primary constraints shown on this slide needed to be considered in designing and conducting human factors applications of sequential experimentation. They also provided a series of guidelines for planning sequential research and selecting, describing, and optimizing factors investigated through sequential experimentation.




The supplemental readings by Han, Williges, and Williges (1997) and Williges, Williges, and Han (1993) provide a detailed discussion of techniques, examples, and guidelines for using sequential experimentation in human factors research. The remaining supplemental readings provide details on response surface methodology and the embedded use of sequential research in these procedures.




To summarize the empirical model building techniques described in Section 5, the purpose and characteristics of empirical model building experiments are reviewed, the major components of empirical models are listed, and a sequential experimentation process for human factors research is provided. This topic ends with some overall conclusions, a summary, and a composite list of supplemental readings covered in Section 5.




The purpose of conducting empirical model building research in human factors and ergonomics is to develop prediction equations of human performance in complex systems based on the functional relationships among quantitative interface variables. These experiments need to provide the necessary and sufficient data to solve complete second-order models expressed in terms of polynomial regression. Model building experiments can also be used to interrogate complex data spaces to determine levels of optimal performance through response surface methodology and sequential experimentation. In addition, Middlebrooks and Williges (2002) also described the use of an augmented fractional-factorial design for general interrogation of network simulations through the use of first-order empirical models.

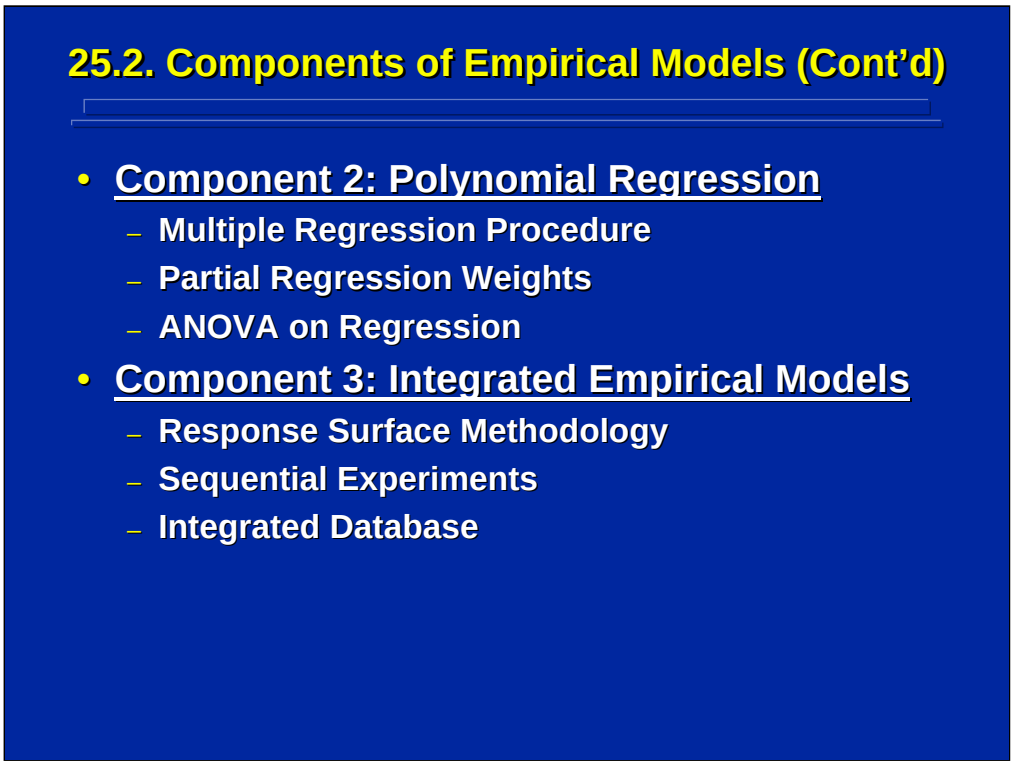
Empirical model building experiments are characterized by the four major characteristics listed at the bottom of this slide. Usually a series of small experiments is used to describe a complex data space. These experiments address meaningful investigations in their own right, use a minimum of replication, and have decision rules for stopping further investigation. These small experiments can be combined into a composite database to build integrated empirical models.

## 25.2. Components of Empirical Models

- **Three Major Components of Empirical Model Building**
  - **Data Collection**
  - **Polynomial Regression**
  - **Integrated Databases**
- **Component 1: First- and Second-Order Experimental Designs**
  - **First-Order Experimental Designs**
    - **$2^k$ Factorial Designs**
    - **Resolution V, $2^{k-p}$ Fractional-Factorial Designs**
  - **Second-Order Experimental Designs**
    - **Central-Composite Designs**

There are three major components of empirical model building. This slide summarizes the first of these three components, and the remaining two components are shown on the next slide.

Component 1 involves the choice of experimental designs used to collect the data needed for building the empirical model. Both first-order and second-order experimental designs are used in building complete second-order empirical models. First-order designs include $2^k$ factorial designs and $2^{k-p}$ fractional factorial designs described in Topic 18. A Resolution V fractional replication is used to keep main effects and two-way interactions orthogonal to each other in building a second-order empirical model. Second-order experimental designs investigate main effects, two-way interactions, and the pure quadratic effects of factors included in the empirical model. Central-composite designs as described in Topic 23 are quite useful in collecting data for complete second-order empirical models because these designs also collect a small amount of additional data that can be used to evaluate lack of fit of the second-order model.




Component 2 deals with using polynomial regression to build second-order empirical models based on the data collected in Component 1. Multiple regression procedures are discussed in Topic 22. Multiple linear regression represents only first-order effects. Polynomial regression is the general form of multiple regression that allows first-, second-, and higher-order effects to be represented in the empirical model. A subsequent ANOVA can be conducted on the polynomial regression to determine the significance of the partial regression weights and the adequacy of fit of the resulting empirical model.

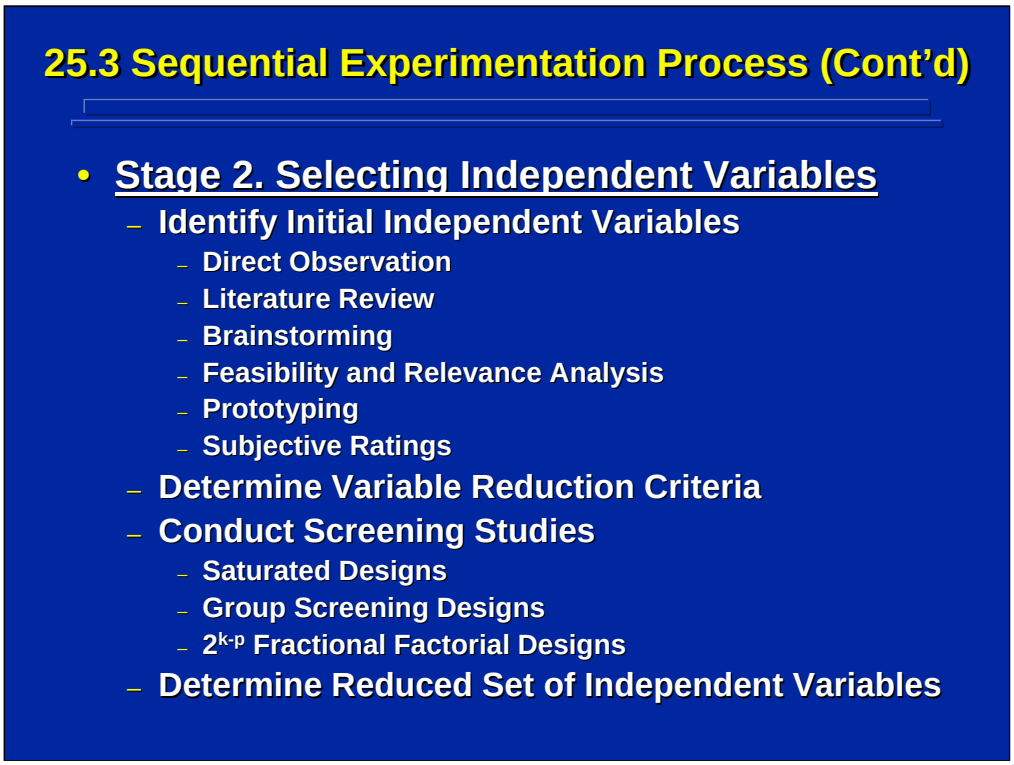
Component 3 provides an extension of empirical model building based on sequential experiments as described in Topic 23. By using techniques of response surface methodology and sequential experimentation, the experimenter can build an integrated database that combines data across a series of small, interrelated experiments. This integrated database can then be used to build integrated empirical models containing more factors than one can investigate in any single experiment in the series.

## 25.3 Sequential Experimentation Process

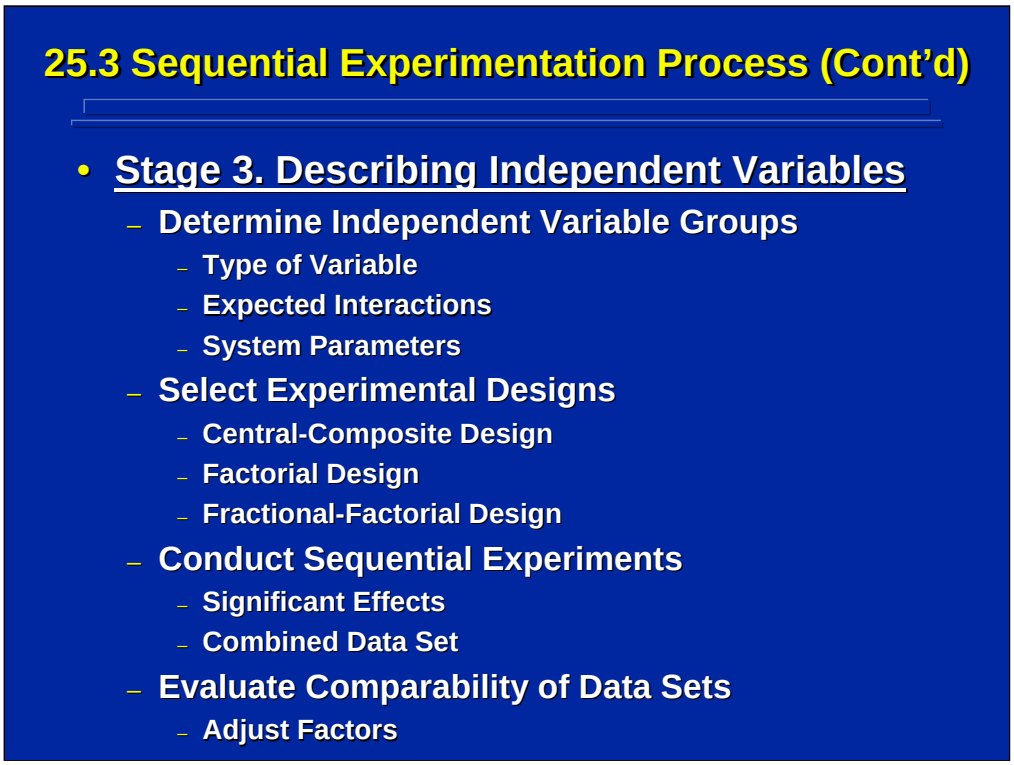
- Han, Williges, and Williges (1997) Paradigm
- Stage 1. Planning Sequential Experimentation
  - Define The System
    - Experimental Region of Interest
    - Assumptions of Empirical Model
  - Determine Experimental Constraints/Requirements
    - Data Collection Strategy
    - Amount of Data
    - Replication
    - Stopping Criteria
    - Benchmark Task
    - Experimental Procedure
    - System Parameter
    - Common Design Point
  - Develop Data Recording Tool

Sequential experimentation as described in Topic 24 is a critical part of the empirical model building repertoire of procedures. Han, Williges, and Williges (1997) provided an extensive flowchart paradigm accompanied by an example of using their paradigm for conducting sequential experimentation in a human factors problem related to a passenger seat design used in a transportation system. Their paradigm lists four major stages with major and minor considerations in each stage. These various considerations are listed in this subsection as an initial guide for the human factors and ergonomics researcher who is considering sequential research while investigating complex systems.

Stage 1 is concerned with the overall plan of sequential experimentation. As shown on this slide, Han, Williges, and Williges (1997) recommend three major planning considerations dealing with defining the boundaries of the experimental space of interest, determining experimental procedural constraints and requirements, and developing a comprehensive recording tool that can be used across the series of experiments for investigating all the factors of interest.

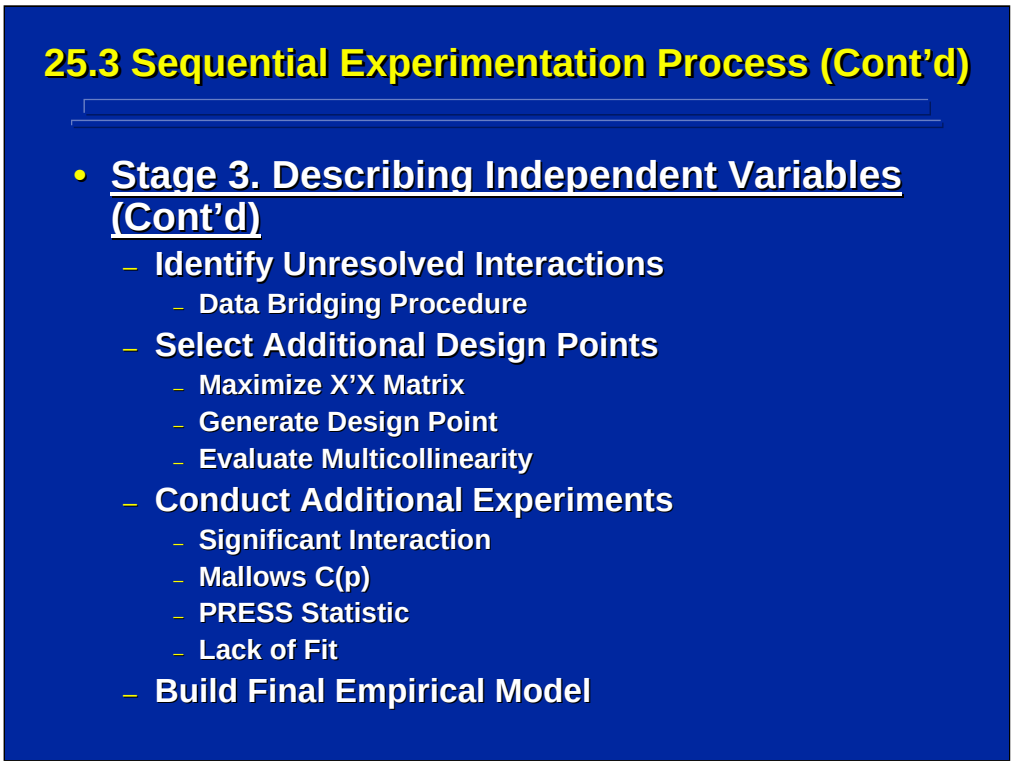



Major and minor considerations for Stage 2 of the Han, Williges, and Williges (1997) paradigm are shown on this slide. A variety of non-experimental and experimental techniques are useful during the four major considerations including the initial identification of potential variables of interest, stating criteria for reducing the initial set of potential factors, conducting screening studies to further reduce the number of factors to be investigated, and the final determination of the factors to investigate through sequential experimentation.

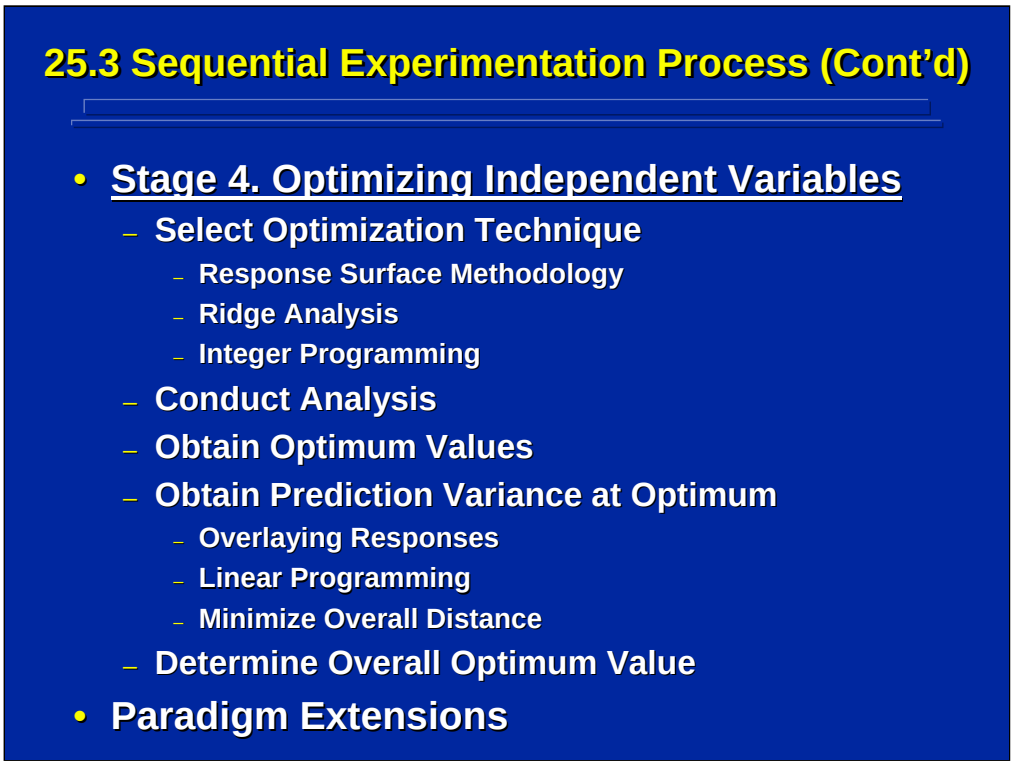



Stage 3 of the Han, Williges, and Williges (1997) paradigm is the major data collection and analysis phase of their sequential experimentation approach. Four of the eight major considerations in their paradigm flowchart are listed on this slide along with minor considerations within each major topic.

Major variables to be investigated need to be grouped into meaningful subsets forming a series of interrelated experiments for subsequent sequential experimentation. These subsets of variables are investigated in small, related experiments using economical experimental designs. The series of experiments is conducted in a meaningful sequence to allow integration into a combined database. Comparability of data across experiments is assessed by evaluating common data points across experiments, and factor adjustments are made, if necessary.




The final four major considerations in Stage 3 of the Han, Williges, and Williges (1997) flowchart are listed on this slide. All four of these considerations are related to data bridging across previous experiments in sequential research series in order to build an integrated database for generating second-order models. An integrated empirical model based on the integrated database including all experiments in the sequence and data bridging runs is the final major consideration in the Stage 3 description phase.




Stage 4 of the Han, Williges, and Williges (1997) paradigm is the final stage and deals with optimizing the interface design. Five major considerations are listed on this slide for interface optimization using empirical models resulting from the integrated database in Step 3.

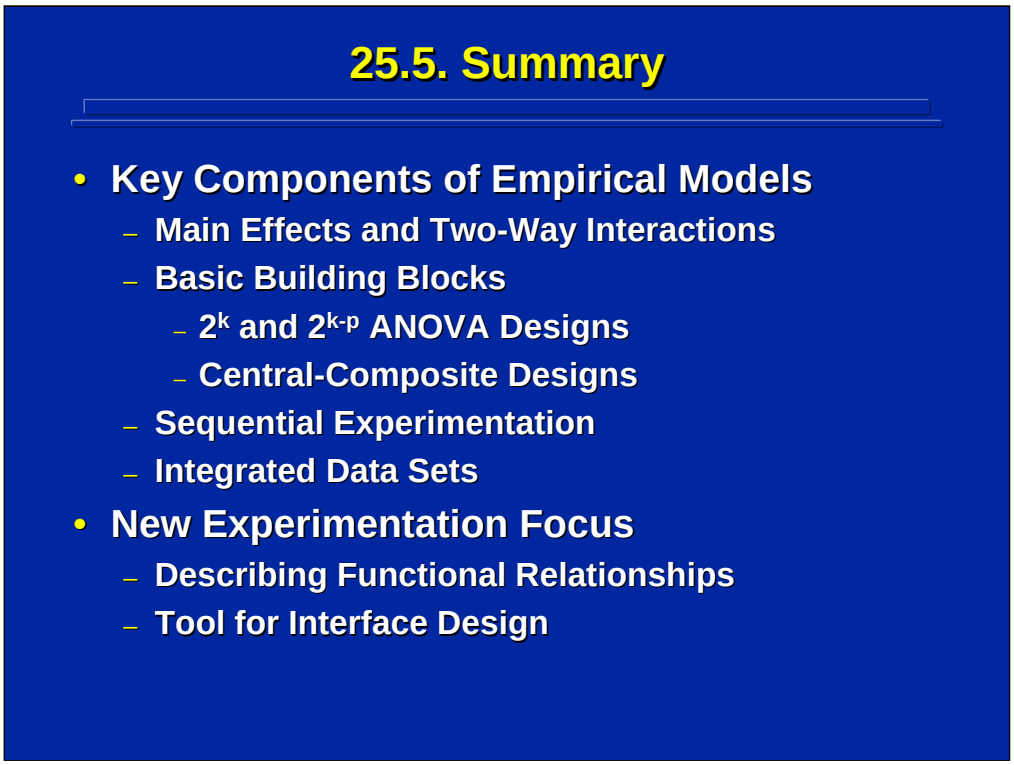
The Han, Williges, and Williges (1997) paradigm should be used in conjunction with the Williges, Williges, and Han (1993) guidelines as discussed in Topic 24 for conducting sequential experimentation. As human factors and ergonomic researchers have more experience with using sequential research procedures, this initial paradigm and guidelines for sequential experimentation will need to be expanded.

## 25.4. Overall Conclusions

- Experimental Design and Analysis Reference
  - Applied Human Factors and Ergonomics Research
    - Sequential Research on Complex Problems
    - Toolkit of Integrated Procedures
  - Improved Experimentation and Methodology
- Computer-Assisted Tool Requirements
  - Interactive Use on Desktop Computers
    - Linked to Facilitate Access
    - Hyperlinked to Interactive Statistical Software
  - Experimental Design Procedures and Processes
  - Reference to Statistical Literature

In terms of overall conclusions, this experimental design and analysis reference material is focused on applied human factors and ergonomics research in complex systems. It is fitting that this reference material ends with sequential experimentation because it underscores the nature of human factors research. Often more than a single experiment is required to investigate interface problems in complex systems. Sequential experimentation requires that the researcher has knowledge of a variety of basic and advanced experimental design and analysis techniques that can be combined to investigate applied, real-world problems. New and improved experimental design procedures, however, are still needed to investigate these complex problems.

This reference material is best provided as a computer-based tool resident on the human factors researcher's desktop computer. The complexity of the material requires extensive linking to facilitate rapid access and review. In addition, hyper-linking the reference material to statistical packages such as SAS (2004) allows the researcher immediate access to interactive statistical analysis of data and a better understanding of these various procedures. This reference material emphasizes the applied research process and uses a building block approach to experimental design to facilitate application and tradeoff of these techniques to complex problems. Finally, the researcher needs to refer to the statistical literature for a deeper understanding of these techniques before choosing to use them.




By way of summary, the top portion of this slide lists four key components of empirical models as discussed in Section 5. The focus of empirical models used in human factors is on predicting the influence of the main effects and two-way interactions on human performance. Second-order empirical models that are needed to describe these effects can be built by using a series of small $2^k$ factorial designs, $2^{k-p}$ fractional replicates, and central-composite designs. These designs can be used in a series of small, sequential experiments. The data can be combined across these interrelated experiments to generate an integrated database that includes many factors of interest in complex research problems that involve more factors than can be feasibly investigated in one large experiment.

Empirical model building provides a new focus for experimentation in human factors and ergonomics research. Rather than just testing the significance of factors, empirical model building investigates functional relationships among factors to predict human performance. These prediction equations, subsequently, can be used for interface design tradeoffs.

## 25.6. Supplemental Readings

| REFERENCE | SECTION |
|---|---|
| Box, Hunter, & Hunter (1978) | Chapters 9, 15-16 |
| Box, Hunter, & Hunter (2005) | Chapters 11, 12 |
| Box and Draper (1987) | Chapters 1-3, 6, 9-12, 14-15 |
| Draper & Smith (1981) | Chapters 2-5 |
| Han, Williges, & Williges (1997) | Entire Article |
| Montgomery (2005) | Chapters 10-11 |
| Myers (1990) | Chapters 3-5, 8, App. A |
| Myers & Montgomery (2002) | Chapters 1-2, 5-8, 14 |
| Wickens (1992) | Chapters 1-2, 7, 9, 11 |
| Williges (1981) | Entire Chapter |
| Williges, Williges, & Han (1993) | Entire Chapter |

This slide provides a summary of supplemental readings on topics presented in Section 5. Empirical Model Building. Han, Williges, and Williges (1997), Williges (1981), and Williges, Williges, and Han (1993) provide details on empirical model building in human factors research. Box and Draper (1987) and Myers and Montgomery (2002) provide a comprehensive coverage in the statistical literature of empirical model building methods. The remaining references deal with specific topics covered in Section 5 dealing with modeling, multiple regression, central-composite designs, and sequential experimentation.

## Acknowledgements

This material is based on over 35 years of teaching graduate courses in applied experimental design and analysis for students in human factors engineering and ergonomics both at Virginia Polytechnic Institute and State University and at the University of Illinois at Urbana-Champaign. Throughout those years, graduate students often expressed their desire to have readily accessible reference material that they could use in the course of designing efficient experiments and in choosing the appropriate data analysis procedures for their human factors research project. Sam Middlebrooks, Chief of the Fort Hood Field Element of the Human Research and Engineering Directorate (HRED) of the U.S. Army Research Laboratory (ARL), and Michael Strub, HRED Manager of the ARL Collaborative Technology Alliance on Advanced Decision Architectures (ADA), provided the impetus to make this happen through this research project. The focus of this project is to provide computer-based reference material on human factors experimental design and analysis in the form of an interactive desktop tool to be used by the human factors researcher. The encouragement of all my former graduate students and the ARL financial support provided by HRED through the ADA prime contract to Micro Analysis and Design are gratefully acknowledged.

Several individuals directly supported this effort by conducting statistical analyses, reviewing draft material, and assisting in the production of the computer media. I want to thank Carrie Slater for conducting SAS analyses, retrieving library references, transcribing notes, and providing comments on the initial draft material. I also thank Beverly Williges for her computer advice/support and for proofreading the final version of this report. Jason Lockhart, the Director of High Performance Computing at Virginia Tech, provided assistance in computer media production. Sam Middlebrooks directed the HRED formative evaluation of draft reference material, and several members of the HRED technical staff including James Ainsworth, Elizabeth Bowman, Otto Heuckeroth, Andrea Krausman, Sam Middlebrooks, Paul Rose, and Ron Spencer provided valuable suggestions during formative evaluation to reduce errors, improve the clarity of presentation, and make the reference material more complete, usable, and interactive. I gratefully acknowledge each of them for their helpful inputs, and I appreciate the comments made by all the Army users to increase the accuracy and usability of this reference material.

## References

Barlow, D.H. and Hersen, M. (1984). *Single-case experimental designs: Strategies for studying behavioral change (2nd ed.)*. New York: Pergamon Press.

Beaudet, D.B. and Williges, R.C. (1988). The role of screening studies in sequential research designs. In *Proceedings of the Human Factors Society 32nd Annual Meeting*, (pp. 1174-1178). Santa Monica, CA: Human Factors Society.

Box, G.E.P. (1954). Some theories on quadratic forms applied to the study of analysis of variance problems. II. Effect of inequality of variance and of correlation between errors in a two-way classification. *Annals of Mathematical Statistics*, **25**, 484-498.

Box, G.E.P. (1957). Evolutionary operation: A method for increasing industrial productivity. *Applied Statistics*, **6**, 81-101.

Box, G.E.P. and Behnken, D.W. (1960). Some new three-level designs for the study of quantitative variables. *Technometrics*, **2**, 455-475.

Box, G.E.P. and Draper, N.R. (1987). *Empirical model building and response surfaces*. New York: Wiley.

Box, G.E.P., Hunter, J.S., and Hunter, W.G. (1978). *Statistics for experimenters: An introduction to design, data analysis, and model building*. New York: Wiley.

Box, G.E.P., Hunter, J.S., and Hunter, W.G. (2005). *Statistics for experimenters: design, innovation, and discovery (2nd ed.)*. New York: Wiley.

Box, G.E.P. and Jenkins, G.M. (1976). *Time series analysis: forecasting and control*. San Francisco: Holden-Day.

Box, G.E.P. and Wilson, K.B. (1951). On the experimental attainment of optimal conditions. *Journal of the Royal Statistical Society, Series B (Methodological)*. **13**, 1-45.

Chin, J.P., Diehl, V.A., and Norman, K.L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of ACM CHI'88*. (pp. 213-218). Washington, D.C.

Cicchetti, D.V. (1972). Extension of multiple-range tests in interaction tables in the analysis of variance: a rapid approximate solution. *Psychological Bulletin*, **77**, 405-408.

Clark, C. and Williges, R.C. (1973). Response surface methodology central-composite design modifications for human performance research. *Human Factors*, **15**, 295-310.

Cochran, W.G. and Cox, G.M. (1957). *Experimental designs (2nd ed.)*. New York: Wiley.

Coleman, W.D., Williges, R.C., and Wixon, D.R. (1985). Collecting detailed user evaluations of software interfaces. *Proceedings of the Human Factors Society 28th Annual Meeting,* (pp. 240-244), Santa Monica: Human Factors Society.

Collier, R.O., Baker, F.B., Manville, G.K., and Hayes, H.T. (1967). Estimates of test size for several test procedures based on variance ratios in the repeated measures design. *Psychometrika*, **32**, 339-353.

Conover, W.J. (1999). *Practical nonparametric statistics (3rd ed.)*. New York: Wiley.

Cook, T.D. and Campbell, D.T. (1979). *Quasi-experimentation design and analysis issues for field settings*. New York: Rand McNally.

Cotton, J.W. (1998). *Analyzing within-subjects experiments*. New Jersey: Lawrence Erlbaum Associates.

Delucchi, K.L. (1993). On the use and misuse of chi-square. In G. Keren and C. Lewis (Eds.) *Data analysis in the behavioral sciences statistical issues*. (pp. 295-320) New Jersey: Lawrence Erlbaum Associates.

Diamond, W.J. (1981). *Practical experimental designs for engineers and scientists*. Belmont, CA: Lifetime Learning.

Draper, N.R. and Smith, H. (1981). *Applied regression analysis (2nd ed.)*. New York: Wiley.

Ericsson, K.A. and Simon, H.A. (1985). *Protocol analysis: Verbal reports as data*. Massachusetts: MIT Press.

Flanagen, J.C. (1954). The critical incident technique. *Psychological Bulletin*, **51**, 327-358.

Geisser, S. and Greenhouse, S.W. (1958). An extension of Box's results on the use of the F distribution in multivariate analysis. *Annals of Mathematical Statistics*, **29**, 885-891.

Gliner, J.A. and Morgan, G.A. (2000). *Research methods in applied settings: an integrated approach to design and analysis*. New Jersey: Lawrence Erlbaum Associates.

Goldstein, I.L. and Ford, J.K. (2002). *Training in organizations needs assessment, development, and evaluation (4th ed.)*. California: Wadsworth/Thompson.

Han, S.H., Williges, B.H., and Williges, R.C. (1997). A paradigm for sequential experimentation. *Ergonomics*, **7**, 737-760.

Hartson, H.R., Andre, T.S., and Williges, R.C. (2003). Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction*, **15**(1), 145-181.

Hays, W.L. (1994). *Statistics (5th ed.)*. California: Wadsworth/Thomson.

Hays, W.L. and Winkler, R.L. (1971). *Statistics: probability, inference, and decision*. New York: Holt, Rinehart & Winston

Heuckeroth, O.H. (2004). Personal communication.

Hicks, C.R. and Turner, K.V. (1999). *Fundamental concepts in the design of experiments (5th ed.)*. New York: Oxford.

Huynh, H. and Feldt, L.S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, **1**, 69-82.

Imhof, J.P. (1962). Testing the hypothesis of no fixed main-effects to Scheffe's mixed model. *Annals of Mathematical Statistics*, **33**, 1085-1095.

Keppel, G. and Wickens, T.D. (2004). *Design and analysis: A researcher's handbook*. *(4th ed.)*. New Jersey: Prentice Hall.

Kies, J.K., Williges, R.C., and Rosson, M.B. (1998). Coordinating computer-supported cooperative work: a review of research issues and strategies. *Journal of the American Society for Information Science,* **49**(9), 776-791.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, **140**, 5-54.

Lindquist, E.F. (1956). *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin.

Martin, D.W. (2004). *Doing psychology experiments (6th ed.)*. California: Wadsworth/Thomson.

Mason, R.L., Gunst, R.F., and Hess, J.L. (2003). *Statistical design and analysis of experiments (2nd ed.)*. New York: Wiley.

Maxwell, S.E. and Delaney, H.D. (2000). *Designing experiments and analyzing data.* New Jersey: Lawrence Erlbaum Associates.

Meister, D. (1985). *Behavioral analysis and measurement methods*. New York: Wiley.

Merkle, P.J., Beaudet, D.B., Williges, R.C., Herlong, D.W., and Williges, B.H. (1988). A systematic method for the selection of independent variables in the investigation of complex systems. In *Proceedings of the Human Factors Society 32nd Annual Meeting*, (pp.1179-1182). Santa Monica, CA: Human Factors Society.

Middlebrooks, S.E. and Williges, R.C. (2002). Experimental design interrogation of network simulation models of U.S. Army command and control centers. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*, (pp. 458-462), Santa Monica: Human Factors and Ergonomics Society.

Mills, R.G. (1977). Personal communication.

Mills, R.G. and Williges, R.C. (1973). Performance prediction in a single-operator simulated surveillance system. *Human Factors*, **15**, 327-348.

Montgomery, D.C. (2005). *Design and analysis of experiments (6th ed.)*. New York: Wiley.

Myers, J.L. (1979). *Fundamentals of experimental design (3rd ed.)*. Boston: Allyn and Bacon.

Myers, J.L. and Well, A.D. (2003). *Research design and statistical analysis (2nd ed.)*. New Jersey: Lawrence Erlbaum Associates.

Myers, R.H. (1990). *Classical and modern regression with applications (2nd ed.)*. Boston: PWS-Kent.

Myers, R.H. and Montgomery, D.C. (2002). *Response surface methodology (2nd ed.)*. New York: Wiley.

Norton, D.W. (1952). An empirical investigation of the effects of non-normality and heterogeneity upon the F-test of analysis of variance. Unpublished Ph.D. dissertation. Iowa State University.

Nunnally, J.C. (1967). *Psychometric theory*. New York: McGraw-Hill.

Osgood, C.E. (1962). Studies on the generality of affective meaning systems. *American Psychologist*, **17**, 10-28.

Pedhazur, E.J. (1982). *Multiple regression in behavioral research (2nd ed.)*. New York: Holt, Rinehart, and Winston.

Pew, R.W. (1993). *Experimental design methodology assessment*. BBN Report No. 7917, Cambridge: Bolt Beranek & Newman, Inc.

Post, D.L. (1981). Personal communication.

Poulton, E.C. (1969). Bias in experimental comparisons between equipment due to the order of testing. *Ergonomics*, **12**, 675-687.

Roscoe, S.N. and Williges, R.C. (1975). Motion relationships in aircraft attitude and guidance displays: A flight experiment. *Human Factors*, **17**, 374-387.

SAS Institute (2004). SAS system version 9.1.3 for Windows and online documentation. North Carolina: The SAS Institute.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, **2**, 110-114.

Schuler, D. and Namioka, A. (1993). *Participatory design: principles and practices*. New Jersey: Lawrence Erlbaum Associates.

Shadish, W.R., Cook, T.D., and Campbell, D.T., (2002). *Experimental and quasi-experimental designs for general causal inference*. New York: Houghton Mifflin.

Siegel, S. and Castellan, N.J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

Simon, C.W. (1977a). Design, analysis, and interpretation of screening studies for human factors engineering research. Westlake Village, CA: Canyon Research Group. Technical Report CWS-03-77.

Simon, C.W. (1977b). New research paradigm for applied experimental psychology: A systems approach. Westlake Village, CA: Canyon Research Group. Technical Report CWS-04-77.

Slater, C.R. and Williges (2006). Appendix: SAS examples for human factors experimental design and analysis reference. Blacksburg, VA: Virginia Polytechnic Institute and State University. Technical Report HFECC-06-02.

Snow, M.P. and Williges, R.C. (1998). Empirical models of perceived presence in virtual environments based on free-modulus magnitude estimation. *Human Factors*, **40**(3), 386-402.

Spine, T.M., Williges, B.H., and Maynard, J.F. (1984). An economical approach to modeling speech recognition accuracy. *International Journal of Man-Machine Studies*, **21**, 191-202.

Stevens, S.S. (1951). Mathematics, measurement, and psychophysics. In S.S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley.

Swain, R.E. (1990). Personal communication.

Tatsuoka, M. (1993). Elements of the general linear model. In G. Keren and C. Lewis (Eds.) *Data analysis in the behavioral sciences statistical issues*. (pp. 3-42) New Jersey: Lawrence Erlbaum Associates.

Tufte, E.R. (1983). The visual display of quantitative information. Connecticut: Graphics Press.

Tufte, E.R. (1990). Envisioning information. Connecticut: Graphics Press.

Tufte, E.R. (1997). Visual exploration: images and quantities, evidence and narrative. Connecticut: Graphics Press.

Walpole, R.E., Myers, R.H., Myers, S.L., and Ye, K. (2002). *Probability and statistics for engineers and scientists* ($7^{th}$ ed.). New Jersey: Prentice Hall.

Wickens, C.D. (1992). *Engineering psychology and human performance ($2^{nd}$ ed.)*. New York: HarperCollins.

Wilk, M.B. and Kempthorne, O. (1957). Nonadditivities in a Latin square design. *Journal of the American Statistical Association*, **52**, 218-236.

Williams, E.J. (1949). Experimental designs balanced for estimation of residual effects of treatments. *Australian Journal of Scientific Research,* **A2***,* 149-168.

Williams, E.J. (1950). Experimental designs balanced for pairs of residual effects. *Australian Journal of Scientific Research*, **A3**, 351-363.

Williges, R.C. (1969). Within-session criterion changes compared to an ideal observer criterion in a visual monitoring task. *Journal of Experimental Psychology,* **81**, 61-66.

Williges, R.C. (1981). Development and use of research methodologies for complex system/simulation experimentation. In J. Moraal and K.F. Kraiss (Eds.) *Manned systems design, methods, equipment, and applications,* (pp. 59-87). New York: Plenum Press.

Williges, R.C. (1987). The use of models in human-computer interface design. *Ergonomics*, **30**, 491-502.

Williges, R.C. (1995). Review of experimental design. In J. Weimer (Ed.), *Research techniques in human engineering.* (pp. 49-71) New Jersey: Prentice Hall PTR.

Williges, R.C. (2006). Response surface methodology and sequential experimentation. In W. Karwowski (Ed.), *International encyclopedia of ergonomics and human factors (2nd ed.)*. New York: Taylor & Francis Ltd.

Williges, R.C. and Baron, C.W. (1973). Transfer assessment using a between-subjects central-composite design. *Human Factors*, **15**, 311-320.

Williges, R.C., Johnston, W.A., and Briggs, G.E. (1966). Role of verbal communication in teamwork. *Journal of Applied Psychology,* **50**, 473-478.

Williges, R.C. and North, R.A. (1973). Prediction and cross-validation of video cartographic symbol location performance. *Human Factors*, **15**, 321-326.

Williges, R.C. and Mills, R.G. (1973). Predicted validity of central-composite design regression equations. *Human Factors*, **15**, 349-354.

Williges, R.C. and Simon, C.W. (1971). Applying response surface methodology to problems of target acquisition. *Human Factors*, 13(**6**), 511-519.

Williges, R.C. and Williges, B.H. (1982). Modeling the human operator in computer-based data entry. *Human Factors*, 24(**3**), 285-299.

Williges, R.C. and Williges, B.H. (1989). Integrated research paradigm for complex experimentation. In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 606-610). Santa Monica, CA: Human Factors Society.

Williges, R.C., Williges, B.H., and Han, S.H. (1992). Developing quantitative guidelines using integrated data from sequential experiments. *Human Factors*, 34(**4**), 399-408.

Williges, R.C., Williges, B.H., and Han, S.H. (1993). Sequential experimentation in human-computer interface design. In R. Hartson and D. Hix (Eds.), *Volume IV, Advances in human-computer interaction*. (pp. 1-30) New York: Ablex.

Winer, B.J., Brown, D.R. and Michels, K.M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

## Subject Index

Additive law, 75
Additivity assumption, 599-600
Adjusted $R^2$, 731
Advanced ANOVA design process, 696
Advantages of ANOVA designs, 242
Algorithm for stating E(MS), 289-296
Aliases, 562
Alpha error, 113-114
Alternative hypothesis, 111
ANCOVA adjusted treatment means, 681
ANCOVA advanced procedures, 683
ANCOVA assumptions, 671
ANCOVA calculations, 677-680
ANCOVA comparisons of adjusted means, 682
ANCOVA interpretation, 684
ANCOVA procedure, 614
ANCOVA readings, 686
ANOVA degrees of freedom, 280-284
ANOVA design classification, 467-470
ANOVA design construction, 308
ANOVA design notation, 317
ANOVA design process, 472-474
ANOVA design, n-factor, between-subjects, 339
ANOVA design, n-factor, mixed-factors design, 455
ANOVA design, n-factor, within-subjects, 416
ANOVA design, one-factor, between-subjects, 312-323
ANOVA design, one-factor, within-subjects, 403-409
ANOVA design, three-factor, between-subjects, 337-338
ANOVA design, three-factor, mixed-factors design, 453-454
ANOVA design, two-factor, between-subjects, 324-335
ANOVA design, two-factor, mixed-factors design, 444-452
ANOVA design, two-factor, within-subjects, 410-415
ANOVA F-tests 297-298
ANOVA hypothesis test, 270-271
ANOVA interaction, 327
ANOVA mean squares, 286
ANOVA n-factor generalizations, 471
ANOVA rules and algorithms, 461-466
ANOVA sources of variation, 276-279
ANOVA statistical models, 250-257
ANOVA sum of squares, 285
ANOVA summary table readings, 310
ANOVA summary table, 304-307

Backward selection procedure, 739
Balanced Latin Square design, 428-436
Balanced sets of treatments, 523-526
Balancing order of treatments, 423-427
Baseline design, 56
Basic ANOVA readings, 273
Basic statistics readings, 140
Beta error, 113-114
Between-subjects ANOVA designs readings, 341
Between-subjects design, 246-247
Binomial distribution, 85-87
Bipolar adjective scales, 167
Blocking $2^k$ ANOVA designs, 533-553
Blocking design considerations, 558
Blocking design readings, 559
Bonferroni t test (Dunn test), 367
Box-Behnken design, 806-808

Case study, 54
CCD background, 767
CCD definition, 772
CCD readings, 810
CCD replication, 775
CCD value of $\alpha$, 776-784
CCD, between-subjects alternative, 788-789
CCD, between-subjects example, 796-801
CCD, blocked, 777-778
CCD, cuboidal, 781
CCD, mixed-factors alternative, 792-793
CCD, orthogonal, 779
CCD, rotatable, 776
CCD, spherical, 780
CCD, within-subjects alternative, 790-791
CCD, within-subjects example, 802-805
Cell, 243-245
Central limit theorem, 88
Chi-square distribution, 94-95
Chi-square test of independence, 185-194
Choice of multiple comparison procedure, 374
Circularity, 417
Cochran and Cox t', 129
Cochran Q test, 199-201
Coefficient of Determination, 669
Cohort designs, 61
Combinations, 79
Complement, 76

Complete hierarchical designs, 494 & 504-508
Complex blocking of $2^k$ designs, 539-542 & 550-553
Complex comparison, 346
Component SS formulae, 527-530
Compositional techniques, 74-76
Conditional probability, 76
Confidence intervals, 103-107
Contingency table (2x2), 186-188
Contingency table (RxC), 191
Control of nuisance factors, 691
Control of subject variability, 694-695
Correlation, covariance formula, 621
Correlation, deviation score formula, 621
Correlation, raw score formula, 622
Correlation, t-test of significance, 628
Correlation, Z-test between two correlations, 631-633
Correlation, Z-test of population value, 630
Counting techniques, 77-80
Critical difference, 358
Critical incidents, 159
Crossed factors, 243-245

Data bridging, 836-837 & 861
Data collection limitations, 692-693
Defining relationship, 535
Dependent variable, 25
Design effects, 580-583
Design efficiency, 590-591
Design resolution, 584-585
Deviation scores, 268
Difference score formula, 135, 144-145
Differential transfer, 437
Dunnett Test, 370

Empirical model components, 856-857
Empirical models, 704, 707
Empirical probability, 73
Estimation of population variance, 269
Estimators, 100-101
Evaluation of interactions readings, 400
Expected mean squares, 287-296
Expected value, 100
Experimental design readings, 70
External validity, 49

F distribution, 96-98

Factor level, 243-245
Factor, 243-245
Factorial design, 243-245
Fisher Z transformation, 629
Fixed-effects variable, 288
Forward selection procedure, 740
Fractional replicates, 562
Fractional-factorial design computations, 572-578
Fractional-factorial design readings, 612
Fractional-factorial design, one-fourth replicate, 569-571
Fractional-factorial design, one-half replicate, 565-568
Fractional-factorial designs, 560
Friedman two-way ANOVA, 222-225
F-test assumptions, 263-265
F-test format, 262
F-test hypotheses, 261

Geisser-Greenhouse correction, 419-422
Generalized interaction, 539
Goodman chi-square, 209
Goodness of fit test, chi-square, 183-185
Goodness of fit test, Kolmogorov-Smirnov, 208
Graphic rating scales, 165
Graphing procedures, 381-383
Greco-Latin square design, 610

Happenstance data, 719-720
Hartley $F_{max}$ test, 128
Hierarchical design definitions, 493-495
Hierarchical design examples, 503-515
Hierarchical design readings, 517
Hierarchical design, between-subjects, 497-498
Hierarchical design, mixed-factors, 501-502
Hierarchical design, within-subjects, 499-500
Homogeneity of covariance, 417-422
Homogeneity of variance, 128
Huynh-Feldt correction, 419-422
Hypothesis test format, 112
Hypothesis testing components, 110

Identity relationship, 562
Independence, 76
Independent variable, 25
Inflated Type I error, 349-350
Institutional Review Board (IRB), 27
Integrated research database, 849-851

Interaction evaluation process, 398
Interaction, 327
Internal validity, 49
Interrupted time series designs, 63-66
Intersection, 74
Interval estimation, 103-107
Interval scale, 150-151

Joint probability, 75

Kolmogorov-Smirnov tests, 207-211
Kruskal-Wallis one-way ANOVA, 212-216

Latin square design definition, 593
Latin square design examples, 606-609
Latin square design readings, 612
Latin square design, balanced, 596
Latin square design, between-subjects, 601-602
Latin square design, standard, 595
Latin square design, within-subjects, 603-604
Laws of probability, 75
Least Significant Difference (LSD) test, 366
Least squares criterion, 101
Level of confidence, 113-114
Level of significance, 113-114
Likert rating scales, 166
Linear comparisons, 345-348
Linear correlation, 616
Literature review, 22

Mallows C(p), 744
McNemar change test, 196-198
Mechanistic models, 704, 706
Method of least squares, 657
Method of steepest ascent, 820-821
Mixed-factors ANOVA design readings, 458
Mixed-factors design, 246 & 249
Model building experiments, 855
Modular arithmetic, 521-522
Multiple comparison readings, 400
Multiple comparisons, 343-344
Multiple linear regression ANOVA, 727-728
Multiple linear regression best equation, 737-752
Multiple linear regression correlational solution, 725
Multiple linear regression example problem, 732 -736
Multiple linear regression matrix solution, 724

Multiple linear regression normal equations, 723
Multiple linear regression shrinkage, 730-731
Multiple linear regression, 721-752
Multiple regression procedures, 716-720
Multiple regression readings, 765
Multiplicative law, 75

Nested factors, 243-245
Nested factors of interest, 690
Newman-Keuls Sequential Range Test, 371-373
Nominal data analysis readings, 203
Nominal data, between-subjects, 181-194
Nominal data, within-subject, 195-201
Nominal scale, 150-151
Nonequivalent control group designs, 57-61
Nonequivalent dependent variables design, 59, 65
Nonparametric procedures, 148
Normal distribution, 88-90
Normative models, 711
Null hypothesis, 111

Ordinal data analysis readings, 227
Ordinal data, between-subjects, 206-216
Ordinal data, within-subject, 217-225
Ordinal scale, 150-151
Orthogonal comparison, 347
Orthogonal interaction components, 531-532

Parameters, 814
Partial correlation, 647-651
Partial hierarchical designs, 495 & 509-515
Partitioning variation, 663
Pearson chi-square statistic, 182
Pearson product-moment correlation, 618
Permutations, 78
Phi correlation, 639-641
Planned comparison F-test, 352-357
Point biserial correlation, 635-638
Point estimation, 102
Polynomial function, 716
Polynomial regression computations, 755
Polynomial regression example, 756-763
Polynomial regression, 717
Polynomial regression, second order, 753
Pooled t-test, 127
Predictive models, 712

PRESS statistic, 743
Pretesting, 29
Pretest-posttest design, 55
Probability, 73
Pseudo-factor blocking, 554-557

Qualitative design guidelines, 842-843
Quantitative design guidelines, 841
Quantitative models readings, 714
Quantitative models, 703
Quasi-experimental designs, 53-66
Quasi-F ratio readings, 491
Quasi-F ratio, 480-482
Questionnaires, 160-162
QUIS scale, 172-176

Random sampling, 81
Random-effects factors, 689
Random-effects variable, 288
Randomized blocks construction, 485
Randomized blocks design readings, 491
Randomized blocks design, 483-488
Randomized experimental designs, 67-68
Rating scale reliability/validity, 168
Ratio scale, 150-151
Regression-discontinuity designs, 62
Research notebook, 42
Research process readings, 44
Research reports, 35-40
Resolution III design, 586
Resolution IV design, 587
Resolution V design, 588
Reverse treatment, pretest/posttest design, 60
Robustness, 125
RSM components, 816
RSM goals, 817
RSM steps, 819
Rules for constructing Balanced Latin Squares, 429-431
Rules for determining ANOVA df, 281-284
Rules for determining F-ratios, 299-302

Sampling distribution, 83-84
Satterthwaite correction, 482
Scales of measurement, 150-151
Scheffé Multiple Contrast Procedure, 368
Self reports, 157

Semantic differential, 167
Semipartial correlation, 652-653
Sequential experimentation process, 858-862
Sequential experimentation readings, 853
Sequential research example, 826-844
Sequential research guidelines, 845-848
Sequential research paradigm, 825
Sequential research, 822-823
Simple blocking of $2^k$ designs, 535-538 & 546-549
Simple effects test, 384-386
Simple linear regression, 654
Simple regression, ANOVA, 664-666
Simple regression, example, 659-660
Simple regression, least squares solution, 658
Simple regression, t-test, 667-668
Spearman correlation, 642-646
Sphericity, 417
SS computational formulae algorithm, 318-321
Stages of research, 21
Standard error between means, 124
Standard error of mean, 102
Standardized regression, 661
Standardized scores, 89
Statistical power, 115
Stepwise selection procedure, 741
Strategies for experimentation, 812-815
Student's t distribution, 91-93
Subjective measures, 154-156
Sum of squares calculations, 316
Supplemental data process, 235-237
Supplemental data readings, 178
Switching replications design, 66

Testing order effects, 432-436
Theoretical F, 260
Theoretical models, 710
Threats to validity, 48-49
Trend analysis, 387-390
t-test assumptions, 125
t-test, between-subjects, 127-133
t-test, single sample, 116-121
t-test, two samples, 122-139
t-test, within-subjects, 134-137
Tukey's Honestly Significant Difference (HSD) Test, 369
Two-level ANOVA design, 266-271
Type I error, 113-114

Type II error, 113-114

Unbiased criterion, 100
Unconfounded comparisons, 393-397
Union, 74
Unit normal distribution, 89-90
Unplanned comparisons, 360-374
Untreated control group designs, 57-58, 64

Verbal protocols, 158

Wilcoxon signed ranks test, 218-221
Within-subjects ANOVA design readings, 441
Within-subjects design, 246 & 248

X variable, 25

**Technical Report HFEEC-06-02**

# Appendix:

# SAS Examples for Human Factors

# Experimental Design and Analysis Reference

Carrie R. Slater and Robert C. Williges

**December 2006**

**Prepared For**

Collaborative Technology Alliance on Advanced Decision Architectures
Micro Analysis and Design
Colorado Springs, Colorado

and

U.S. Army Research Laboratory
Human Research and Engineering Directorate
Aberdeen Proving Ground, Maryland 21005-5425

**Human Factors Engineering and Ergonomics Center**
Grado Department of Industrial and Systems Engineering
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

## Table of Contents


| Contents | Page |
|---|---|
| **Introduction** | 1 |
| **Section 1. Introduction to Experimental Design – Example Problems** | |
| 1. Interval Estimation | 4 |
| 2. Single-Sample t-Test | 6 |
| 3. Between-Subjects t-Test | 9 |
| 4. Within-Subjects t-Test | 12 |
| **Section 2. Supplemental Data Collection and Analysis – Example Problems** | |
| 5. Chi-Square Goodness of Fit Test | 15 |
| 6. Chi-Square Test of Independence (2x2 Contingency Table) | 17 |
| 7. Chi-Square Test of Independence (RxC Contingency Table) | 19 |
| 8. Chi-Square Test of Independence (Two Additive 2x2 Partitions) | 22 |
| 9. McNemar Change Test | 26 |
| 10. Cochran Q Test | 28 |
| 11. Kolmogorov-Smirnov Tests | 30 |
| 12. Kruskal-Wallis One-Way ANOVA | 33 |
| 13. Wilcoxon Signed Ranks Test | 35 |
| 14. Friedman Two-Way ANOVA | 38 |
| **Section 3. Basic Analysis of Variance (ANOVA) Designs – Example Problems** | |
| 15. One-Factor, Between-Subjects ANOVA | 41 |
| 16. Two-Factor, Between-Subjects ANOVA | 44 |
| 17. Planned Comparisons | 48 |
| 18. Unplanned Comparisons | 52 |
| 19. Analysis of Interactions | 61 |
| 20. One-Factor, Within-Subjects ANOVA | 77 |
| 21. Two-Factor, Within-Subjects ANOVA | 80 |
| 22. Geisser-Greenhouse and Huynh-Feldt Corrections | 84 |
| 23. Testing Order Effects in Balanced Latin Squares | 94 |
| 24. Within-Subjects and Between-Subjects Design Comparison | 98 |
| 25. Two-Way, Mixed-Factors ANOVA | 103 |
| **Section 4. Advanced ANOVA Designs – Example Problems** | |
| 26. Complete Hierarchical Between-Subjects Design | 108 |
| 27. Partial Hierarchical Between-Subjects Design | 113 |
| 28. Simple Blocking of $2^k$ Within-Subjects Design | 118 |
| 29. Complex Blocking of $2^k$ Within-Subjects Design | 126 |

30. One-Half Replicate of $2^4$ Between-Subjects Design 133
31. 4x4 Latin Square Designs 137
32. Linear Correlation Coefficient 146
33. Alternative Linear Correlations 150
34. Simple Linear Regression 158
35. One-Way, Analysis of Covariance (ANCOVA) 162

**Section 5. Empirical Model Building – Example Problems**

36. Multiple Linear Regression 169
37. Best Regression Equation 172
38. Polynomial Regression 190
39. Orthogonal, Between-Subjects, Central-Composite Design 199
40. Blocked, Within-Subjects, Central-Composite Design 209

**Acknowledgements** 216

**References** 217

## Introduction

This report provides examples of statistical analyses using Version 9.1.3 of the SAS statistical package developed by the SAS Institute (2004). These analyses follow the examples provided by Williges (2006) in his reference material that describes applied experimental design and analysis useful in human factors and ergonomics research. Hyperlinks to the Williges (2006) PDF document are provided throughout this report.

The purpose of this document is to aid users of the Williges (2006) reference material in using a statistical analysis package. Although SAS is used as an example computer package for statistical analysis, many statistical packages are available with similar features and abilities. This document is not intended to provide a detailed discussion of the entire SAS package. The reader is referred to the SAS Institute (2004) online user manual for Version 9.1.3 and to Cody and Smith (1997) for a detailed discussion of conducting various statistical analyses using SAS.

A consistent format is followed for the presentation of each example. Each example provided in this report is referenced to the appropriate discussion in the Williges (2006) report. Each problem described by Williges (2006) is enhanced by providing the context/purpose and the statistical decision criteria. This problem statement is followed by the actual SAS input file stating the SAS procedures and uses the data set from Williges (2006). Each SAS input file is hyperlinked to the actual SAS program file which will appear in the SAS editor when clicked. This feature is provided for SAS users so that the example is readily available in SAS for interactive use and modification for alternative data sets. The SAS output file of the appropriate statistical analysis is presented. Wherever appropriate, special notes about procedures are provided to aid the user in conducting each example analysis with SAS. Finally, an explanation of the results is provided for each example, and the relevant aspects of the SAS output file that are related to the explanation are marked in boldface type for easy reference.

The basic SAS setup of each problem follows a standard procedure. There are some major components to these basic setups including an options statement, title, data set name, input definition, and the procedure commands. Each statement must end with a semi-colon (;) to designate the completion of a single program statement (Cody and Smith 1997). See the SAS Institute (2004) online user manual for a listing of all procedure statements and options. A typical SAS format using data formatted in columns is as follows:

Description Component
**Options Statement:** Allows system options to be added (e.g., page numbers and centering);
**Title:** Allows inclusion of a meaningful problem title and must be enclosed in single quotation marks;

Data Input Component
**Data Statement:** Names the data set being entered;

**Input Statement:** Names the columns and describes the format of the data for the corresponding data to be entered;
**Lines Statement:** Tells the program that the information to follow will be the actual data to be analyzed;
**Data Input:** Lists in columns the data corresponding to the variable names given in the input statement;

Data Analysis Component
**Procedure Statement:** Tells the program which type of statistical analysis to be conducted and any options associated with that analysis;
**Variable Statement:** Tells the program which variables are used in the statistical analysis specified; and
**Quit Statement:** Tells the program that it has reached the end of the analysis.

# Section 1. Introduction to Experimental Design

## Example 1: Interval Estimation

**(Click in this red rectangle to see the Reference Notes on Example 1.)**

### Problem

Location in Williges (2006) Table of Contents
Section 1, Topic 3. Basic Statistical Concepts, Part 3.4.3. Interval Estimation

Page(s) in Williges (2006) Reference Material: 107

Problem Description
The reaction time (RT) of 6 subjects detecting a signal was measured. The mean RT was .657 seconds, and the standard deviation was .0706 seconds. What is the 95% confidence interval of the true mean RT?

Context/Purpose
Determine the range within which the true mean would be expected to occur 95% of the time.

Statistical Decision Criteria
For small sample sizes and the 95% confidence interval, use the t-tabled values below/above which 0.025 of those cases would be expected to occur for the available degrees of freedom.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 1: Interval Estimation';
data Reaction;
input Time;
lines;
0.56
0.77
0.69
0.62
0.64
0.66
;
proc means alpha=0.05 n mean stddev var stderr clm data=reaction;
var Time;
quit;
```

## SAS Output

```
Example 1: Interval Estimation

The MEANS Procedure

                              Analysis Variable : Time

                                                                  Lower 95%       Upper 95%
N            Mean         Std Dev        Variance       Std Error CL for Mean     CL for Mean
------------------------------------------------------------------------------------------------
6       0.6566667       0.0706163       0.0049867       0.0288290   0.5825594       0.7307740
------------------------------------------------------------------------------------------------
```

## Output Explanation

95% of the time the true mean falls between 0.583 and 0.731 seconds.

## Example 2: Single-Sample t-Test

**(Click in this red rectangle to see the Reference Notes on Example 2.)**

### Problem

Location in Williges (2006) Table of Contents
Section 1, Topic 3. Basic Statistical Concepts, Part 3.5.2. Single-Sample t-Test

Page(s) in Williges (2006) Reference Material: 117 – 121

Problem Description
The experimenter wishes to compare the average scores on the final examination in a military course to a standard population value of 792 points. Forty-nine students are randomly assigned to a particular section of the course, and they scored an average of 827.61 points with a standard deviation of 84.19 points. The experimenter is interested in determining if the 827.61 point average is significantly different from the standard value of 792 points. This test is conducted at the 0.05 level of significance.

Context/Purpose
Determine if there is significant difference between a sample mean and a standard mean.

Statistical Decision Criterion
To be conservative, use a two-tailed, t-test to determine if there is a difference between the mean of the class and the known value. Since this is a two-tailed test, $\alpha$ is set at 0.025 when determining the tabled value in a standard t table.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 2: Single Sample t-Test';
data final;
input Scores;
lines;
881
786
665
783
766
998
954
906
763
827
862
793
806
838
874
923
```

```
762
958
686
841
750
805
863
832
678
936
791
812
887
765
816
723
730
868
843
956
934
902
825
884
876
708
721
739
811
792
723
930
981
;
proc ttest h0=792 alpha=0.05 data=final;
var Scores;
quit;
```

## SAS Output

```
Example 2: Single Sample t-Test

The TTEST Procedure
                                        Statistics

                  Lower CL        Upper CL  Lower CL             Upper CL
Variable    N     Mean      Mean     Mean    Std Dev   Std Dev    Std Dev  Std Err

Scores     49    803.43    827.61   851.79   70.207    84.189     105.18  12.027

                T-Tests

Variable       DF     t Value     Pr > |t|

Scores         48        2.96       0.0048
```

## Output Explanation

Since the two-tailed p-value (Pr > |t|) resulting from the SAS analysis (0.0048) is less than 0.05, one can reject the null hypothesis. Therefore the class mean (827.61) is significantly larger than the known population mean (792). Consequently, the population mean, 792, falls outside the 95% confidence interval of the class mean, C[803.43 ≤ $\mu$ ≤ 851.79] = .95.

## Example 3: Between-Subjects t-Test

### (Click in this red rectangle to see the Reference Notes on Example 3.)

**Problem**

Location in Williges (2006) Table of Contents
Section 1, Topic 3. Basic Statistical Concepts, Part 3.6.4. Between-Subjects t-Test

Page(s) in Williges (2006) Reference Material: 130-133

Problem Description
An experimenter wishes to compare performance of two different night vision displays used in nighttime maneuvering. Eight squads used Display A, and eight different squads used Display B. Each squad completed the same nighttime maneuver. The experimenter wants to determine if there is a significant difference ($p < 0.05$) in mean time in minutes to complete the nighttime maneuver using the two night vision displays.

Context/Purpose
Determine if there is a significant difference in the average time for a squad to complete the nighttime maneuver using the two night vision displays.

Statistical Decision Criteria
A two-tailed, between-subjects, pooled t-test conducted at the 0.5 level of significance is appropriate. Since this is a two-tailed test, $\alpha$ is set at 0.025 when determining the tabled value in a standard t table. A preliminary test of homogeneity of variance is usually not needed since sample sizes are equal.

**SAS Input**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options pageno=1 nodate nocenter;
Title 'Example 3: Between-Subjects t-Test';
data display;
input type $ Time;
lines;
A 59
A 65
A 52
A 45
A 63
A 42
A 53
A 47
B 54
B 72
B 69
B 59
B 67
B 61
B 51
```

```
B 63
;
proc means data=display n mean std stderr;
var Time;
by type;
proc ttest ci=equal alpha=0.05 data=display;
class type;
var Time;
quit;
```

## SAS Output

```
Example 3: Between-Subjects t-Test

type=A

The MEANS Procedure

             Analysis Variable : Time

 N            Mean         Std Dev       Std Error
------------------------------------------------------
 8      53.2500000       8.4642104       2.9925503
------------------------------------------------------

type=B

             Analysis Variable : Time

 N            Mean         Std Dev       Std Error
------------------------------------------------------
 8      62.0000000       7.2702918       2.5704363
------------------------------------------------------

The TTEST Procedure

                                    Statistics

                          Lower CL          Upper CL  Lower CL           Upper CL
Variable  type           N      Mean    Mean      Mean   Std Dev  Std Dev  Std Dev  Std Err

Time      A              8    46.174   53.25    60.326    5.5963   8.4642   17.227   2.9926
Time      B              8    55.922      62    68.078    4.8069   7.2703   14.797   2.5704
Time      Diff (1-2)          -17.21   -8.75    -0.289    5.7764   7.8899   12.443   3.9449

                             T-Tests

Variable    Method           Variances      DF    t Value    Pr > |t|

Time        Pooled           Equal          14      -2.22      0.0436
Time        Satterthwaite    Unequal      13.7      -2.22      0.0440

                    Equality of Variances

Variable    Method      Num DF    Den DF    F Value    Pr > F

Time        Folded F         7         7       1.36    0.6984
```

**Output Explanation**

The SAS program automatically tests for homogeneity of variance even if sample size is equal in the two samples. One usually sets a high $\alpha$ error (e.g. $\alpha$ = 0.20) when making a homogeneity of variance test to avoid a Type II error. Since the $F_{max}$ is 1.36 and is significant at $\alpha$ = 0.6984, the experimenter fails to reject the null hypothesis and is justified in assuming homogeneity of variance for the subsequent pooled t-test of difference between using two night vision displays in nighttime maneuvering.

The SAS program conducts a two-tailed t-test and states the probability of $\alpha$ error accordingly (Pr > |t|). Since the two-tailed p-value resulting from the SAS analysis (0.0436) is less than 0.05, one can reject the null hypothesis of equal means. Therefore, using night vision Display A resulted in an average of 8.75 significantly shorter nighttime maneuvering minutes than using night vision Display B (i.e., an average of 53.25 minutes using Display A versus 62 minutes using Display B).

## Example 4: Within-Subjects t-Test

### (Click in this red rectangle to see the Reference Notes on Example 4.)

**Problem**

Location in Williges (2006) Table of Contents
Section 1, Topic 3. Basic Statistical Concepts, Part 3.6.5. Within-Subjects t-Test

Page(s) in Williges (2006) Reference Material: 136-137

Problem Description
An experimenter wishes to compare performance of two different night vision displays used in nighttime maneuvering. Eight squads used both Display A and Display B. Each squad completed the same nighttime maneuver twice. Half of the squads used Display A first and half used Display B first to counterbalance order of use. The experimenter wants to determine if there is a significant difference ($p < 0.05$) in mean time in minutes to complete the nighttime maneuver between using the two night vision displays.

Context/Purpose
Determine if there is a significant difference in the average time for a squad to complete the nighttime maneuver using the two night vision displays.

Statistical Decision Criteria
A two-tailed, within-subjects t-test conducted at the 0.5 level of significance using difference scores is appropriate. Since this is a two-tailed test, $\alpha$ is set at 0.025 when determining the tabled value in a standard t table. No preliminary test of homogeneity of variance is required because repeated measures are used.

**SAS Input**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options pageno=1 nodate nocenter;
Title 'Example 4: Within-Subjects t-Test';
data display;
input A B;
lines;
59 54
65 72
52 69
45 59
63 67
42 61
53 51
47 63
;
proc means data=display n mean var std stderr;
var A B;
proc ttest alpha=0.05 data=display;
paired A*B;
quit;
```

## SAS Output

```
Example 4: Within-Subjects t-Test

The MEANS Procedure
```

| Variable | N | **Mean** | Variance | Std Dev | Std Error |
|---|---|---|---|---|---|
| A | 8 | **53.2500000** | 71.6428571 | 8.4642104 | 2.9925503 |
| B | 8 | **62.0000000** | 52.8571429 | 7.2702918 | 2.5704363 |

```
The TTEST Procedure
```

Statistics

| Difference | N | Lower CL Mean | **Mean** | Upper CL Mean | Lower CL Std Dev | Std Dev | Upper CL Std Dev | Std Err |
|---|---|---|---|---|---|---|---|---|
| A - B | 8 | -16.38 | **-8.75** | -1.117 | 6.0365 | 9.13 | 18.582 | 3.2279 |

**T-Tests**

| **Difference** | **DF** | **t Value** | **Pr > \|t\|** |
|---|---|---|---|
| **A - B** | **7** | **-2.71** | **0.0302** |

## Output Explanation

Since the two-tailed p-value (Pr > |t|) resulting from the SAS analysis (0.0302) is less than 0.05, one can reject the null hypothesis of equal means. Therefore, using night vision Display A resulted in an average of 8.75 significantly shorter nighttime maneuvering minutes than using night vision Display B (i.e., 53.25 minutes versus 62 minutes).

# Section 2. Supplemental Data Collection and Analysis

## Example 5: Chi-Square Goodness of Fit Test

**(Click in this red rectangle to see the Reference Notes on Example 5.)**

### Problem

Location in Williges (2006) Table of Contents
Section 2, Topic 5. Analysis of Nominal Data, Part 5.2.1. Chi-Square Goodness of Fit Test

Page(s) in Williges (2006) Reference Material: 184

Problem Description
The relative frequency of the age of automobile drivers in the U.S. is known. A sample of 50 drivers is chosen, and demographic data on age are recorded in six age groupings. Does the age of this sample differ from the distribution of the U.S. population of drivers known to be 0.19, 0.11, 0.15, 0.27, 0.16, and 0.12 in the six age groupings ($p < 0.01$)?

Context/Purpose
Determine if the sampled demographic data are different from the known population of U.S. drivers.

Statistical Decision Criteria
The chi-square goodness of fit test is the appropriate test to compare the observed category frequencies to known population values. The test is made at the 0.01 level of significance.

### SAS Input**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

***Note: The observed frequencies are listed in the data input. The expected frequencies for each category are listed in the "testp" command and must be stated in terms of relative frequencies or proportion of the sample size.*

```
options nodate nocenter pageno=1;
title 'Example 5: Chi-Square Goodness of Fit Test';
data agedata;
input Age $ observed;
lines;
18-25 10
26-35 3
36-45 6
46-55 25
56-65 5
>65   1
;
proc freq data=agedata;
weight observed;
tables Age/nocum testp=(.19 .11 .15 .27 .16 .12)alpha=0.01;
quit;
```

## SAS Output

Example 5: Chi-Square Goodness of Fit Test

The FREQ Procedure

| Age | Frequency | Percent | Test Percent |
|---|---|---|---|
| 18-25 | 10 | 20.00 | 19.00 |
| 26-35 | 3 | 6.00 | 11.00 |
| 36-45 | 6 | 12.00 | 15.00 |
| 46-55 | 25 | 50.00 | 27.00 |
| 56-65 | 5 | 10.00 | 16.00 |
| >65 | 1 | 2.00 | 12.00 |

**Chi-Square Test for Specified Proportions**

| | |
|---|---|
| **Chi-Square** | **16.5506** |
| **DF** | **5** |
| **Pr > ChiSq** | **0.0054** |

Sample Size = 50

## Output Explanation

Since the p-value resulting from the SAS analysis (0.0054) is less than 0.01, one can reject the null hypothesis. Therefore, the age distribution of the sample of drivers in this study is significantly different from the U.S population of drivers.

## Example 6: Chi-Square Test of Independence (2x2 Contingency Table)

**(Click in this red rectangle to see the Reference Notes on Example 6.)**

### Problem

Location in Williges (2006) Table of Contents
Section 2, Topic 5. Analysis of Nominal Data, Part 5.2.2. Chi-Square Test of Independence

Page(s) in Williges (2006) Reference Material: 189 – 190

Problem Description
Every user in a random sample of 80 users classified themselves as either high (Hi) or low (Lo) in computer experience. All users practiced using an experimental text editor for 10 hours and were then asked to state whether they were satisfied (Yes) or not satisfied (No) with the text editor. Is their satisfaction evaluation independent of their computer experience ($p < 0.05$)?

Context/Purpose
Determine if satisfaction with the text editor is dependent upon amount of computer experience of the user.

Statistical Decision Criteria
A chi-squared test of independence of a 2x2 contingency table is appropriate to compare the frequency of satisfied ratings classified into two qualitative groups of computer experience and satisfaction with the text editor. The test is made at the 0.05 level of significance.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 6: Chi-Square Test of Independence (2x2 Contingency Table)';
data computer;
input Experience $ Satisfied $ count;
lines;
Hi Yes 24
Hi No 11
Lo Yes 16
Lo No 29
;
proc freq data=computer;
tables Experience*Satisfied/chisq expected alpha=0.05;
weight count;
quit;
```

## SAS Output

```
Example 6: Chi-Square Test of Independence (2x2 Contingency Table)

The FREQ Procedure

Table of Experience by Satisfied

Experience     Satisfied
```

| Frequency<br>Expected<br>Percent<br>Row Pct<br>Col Pct | No | Yes | Total |
|---|---|---|---|
| Hi | 11<br>17.5<br>13.75<br>31.43<br>27.50 | 24<br>17.5<br>30.00<br>68.57<br>60.00 | 35<br><br>43.75<br><br> |
| Lo | 29<br>22.5<br>36.25<br>64.44<br>72.50 | 16<br>22.5<br>20.00<br>35.56<br>40.00 | 45<br><br>56.25<br><br> |
| Total | 40<br>50.00 | 40<br>50.00 | 80<br>100.00 |

```
The FREQ Procedure

Statistics for Table of Experience by Satisfied
```

| Statistic | DF | Value | Prob |
|---|---|---|---|
| **Chi-Square** | **1** | **8.5841** | **0.0034** |
| Likelihood Ratio Chi-Square | 1 | 8.7558 | 0.0031 |
| Continuity Adj. Chi-Square | 1 | 7.3143 | 0.0068 |
| Mantel-Haenszel Chi-Square | 1 | 8.4768 | 0.0036 |
| Phi Coefficient | | -0.3276 | |
| Contingency Coefficient | | 0.3113 | |
| Cramer's V | | -0.3276 | |

```
Sample Size = 80
```

## Output Explanation

Since the p-value resulting from the SAS analysis (0.0066) is less than 0.05, one can reject the null hypothesis. Therefore, user satisfaction with the experimental text editor was significantly dependent upon the amount of computer experience.

## Example 7: Chi-Square Test of Independence (RxC Contingency Table)

**(Click in this red rectangle to see the Reference Notes on Example 7.)**

### Problem

Location in Williges (2006) Table of Contents
Section 2, Topic 5. Analysis of Nominal Data, Part 5.2.2. Chi-Square Test of Independence

Page(s) in Williges (2006) Reference Material: 192

Problem Description
Every user in a random sample of 80 users classified themselves as high (Hi), medium (Med) or low (Lo) in computer experience. All users practiced using an experimental text editor for 10 hours and were then asked to state whether they were satisfied (Yes) or not satisfied (No) with the text editor. Is their satisfaction evaluation independent of their computer experience ($p < 0.05$)?

Context/Purpose
Determine if satisfaction with two different text editors is dependent upon amount of computer experience of the users.

Statistical Decision Criteria
Since n is greater than twenty and E is greater than 5, use a chi-square test of independence at the 0.05 level of significance using 3x2 contingency tables to compare frequencies organized in multiple qualitative groups.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 7: Chi-Square Test of Independence (RxC Contingency Table)';
data computer;
input Experience $ Satisfied $ count;
lines;
Hi Yes 24
Hi No 10
Med Yes 8
Med No 7
Lo Yes 8
Lo No 23
;
proc freq data=computer;
tables Experience*Satisfied/chisq expected alpha = .05;
weight count;
quit;
```

## SAS Output

Example 7: Chi-Square Test of Independence (RxC Contingency Table)

The FREQ Procedure

Table of Experience by Satisfied

Experience  Satisfied

| Frequency<br>Expected<br>Percent<br>Row Pct<br>Col Pct | No | Yes | Total |
|---|---|---|---|
| Hi | 10<br>17<br>12.50<br>29.41<br>25.00 | 24<br>17<br>30.00<br>70.59<br>60.00 | 34<br><br>42.50<br><br> |
| Lo | 23<br>15.5<br>28.75<br>74.19<br>57.50 | 8<br>15.5<br>10.00<br>25.81<br>20.00 | 31<br><br>38.75<br><br> |
| Med | 7<br>7.5<br>8.75<br>46.67<br>17.50 | 8<br>7.5<br>10.00<br>53.33<br>20.00 | 15<br><br>18.75<br><br> |
| Total | 40<br>50.00 | 40<br>50.00 | 80<br>100.00 |

Statistics for Table of Experience by Satisfied

| Statistic | DF | Value | Prob |
|---|---|---|---|
| **Chi-Square** | **2** | **13.0894** | **0.0014** |
| Likelihood Ratio Chi-Square | 2 | 13.5782 | 0.0011 |
| Mantel-Haenszel Chi-Square | 1 | 3.7513 | 0.0528 |
| Phi Coefficient | | 0.4045 | |
| Contingency Coefficient | | 0.3750 | |
| Cramer's V | | 0.4045 | |

Sample Size = 80

### Output Explanation

Since the p-value resulting from the SAS analysis (0.0014) is less than 0.05, one can reject the null hypothesis. Therefore, satisfaction with the experimental text editor is statistically dependent on computer experience. One would need to conduct additional tests to determine the locus of computer experience dependency through a series of additional chi-square tests of independence using meaningful 2x2 partitions of the original 3x2 contingency table.

## Example 8: Chi-Square Test of Independence (Two Additive 2x2 Partitions)

**(Click in this red rectangle to see the Reference Notes on Example 8.)**

### Problem

Location in Williges (2006) Table of Contents
Section 2, Topic 5. Analysis of Nominal Data, Part 5.2.2. Chi-Square Test of Independence

Page(s) in Williges (2006) Reference Material: 193-194

Problem Description
Every user in a random sample of 80 users classified themselves as high (Hi), medium (Med) or low (Lo) in computer experience. All users practiced using an experimental text editor for 10 hours and were then asked to state whether they were satisfied (Yes) or not satisfied (No) with the text editor. Is their satisfaction evaluation independent of their computer experience ($p < 0.05$)?

Context/Purpose
Determine which levels of variables within the 3x2 contingency table of Example 7 are independent of each other. First, only users classified as having either high or medium computer experience are compared to determine if their satisfaction evaluation is independent of their computer experience.

Statistical Decision Criteria
All additional chi-square tests of independence using two additive 2x2 partitions use the same level of significance (i.e., $p < 0.05$) as the overall 3x2 contingency table test in order to determine which qualitative groups have significant effects.

### SAS Input (Part A. 2x2 Table 1)

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 8A: Chi-Square Test of Independence (Two Additive 2x2
Partitions)';
data computer;
input Experience $ Satisfied $ count;
lines;
Hi Yes 24
Hi No 10
Med Yes 8
Med No 7
;
proc freq data=computer;
tables Experience*Satisfied/chisq expected alpha = .05;
weight count;
quit;
```

## SAS Output (Part A. 2x2 Table 1)

```
Example 8A: Chi-Square Test of Independence (Two Additive 2x2 Partitions)

The FREQ Procedure

Table of Experience by Satisfied

Experience     Satisfied
```

| Frequency<br>Expected<br>Percent<br>Row Pct<br>Col Pct | No | Yes | Total |
|---|---|---|---|
| Hi | 10 | 24 | 34 |
| | 11.796 | 22.204 | |
| | 20.41 | 48.98 | 69.39 |
| | 29.41 | 70.59 | |
| | 58.82 | 75.00 | |
| Med | 7 | 8 | 15 |
| | 5.2041 | 9.7959 | |
| | 14.29 | 16.33 | 30.61 |
| | 46.67 | 53.33 | |
| | 41.18 | 25.00 | |
| Total | 17 | 32 | 49 |
| | 34.69 | 65.31 | 100.00 |

```
The FREQ Procedure

Statistics for Table of Experience by Satisfied
```

| Statistic | DF | Value | Prob |
|---|---|---|---|
| **Chi-Square** | **1** | **1.3677** | **0.2422** |
| Likelihood Ratio Chi-Square | 1 | 1.3401 | 0.2470 |
| Continuity Adj. Chi-Square | 1 | 0.7122 | 0.3987 |
| Mantel-Haenszel Chi-Square | 1 | 1.3398 | 0.2471 |
| Phi Coefficient | | -0.1671 | |
| Contingency Coefficient | | 0.1648 | |
| Cramer's V | | -0.1671 | |

```
Sample Size = 49
```

## Output Explanation (Part A. 2x2 Table 1)

Since the p-value resulting from the SAS analysis (0.2422) is greater than 0.05, one cannot reject the null hypothesis. Therefore, user satisfaction of text editors cannot be considered statistically dependent on the high and medium levels of computer experience. Next, one would combine users with high and medium levels into one group and compare them to users with low level computer experience in an additional test of significance using a 2x2 contingency table.

### SAS Input (Part B. 2x2 Table 2)

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 8B: Chi-Square Test of Independence (Two Additive 2x2
Partitions)';
data computer;
input Experience $ Satisfied $ count;
lines;
Hi+Med Yes 32
Hi+Med No 17
Lo Yes 8
Lo No 23
;
proc freq data=computer;
tables Experience*Satisfied/chisq expected alpha = .05;
weight count;
quit;
```

### SAS Output (Part B. Table 2)

```
Example 8: Chi-Square Test of Independence (Two Additive 2x2 Partitions)

The FREQ Procedure

Table of Experience by Satisfied

Experience     Satisfied
```

| Frequency<br>Expected<br>Percent<br>Row Pct<br>Col Pct | No | Yes | Total |
|---|---|---|---|
| Hi+Med | 17 | 32 | 49 |
| | 24.5 | 24.5 | |
| | 21.25 | 40.00 | 61.25 |
| | 34.69 | 65.31 | |
| | 42.50 | 80.00 | |
| Lo | 23 | 8 | 31 |
| | 15.5 | 15.5 | |
| | 28.75 | 10.00 | 38.75 |
| | 74.19 | 25.81 | |
| | 57.50 | 20.00 | |
| Total | 40 | 40 | 80 |
| | 50.00 | 50.00 | 100.00 |

```
The FREQ Procedure

Statistics for Table of Experience by Satisfied

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1     11.8499    0.0006
Likelihood Ratio Chi-Square    1     12.2381    0.0005
Continuity Adj. Chi-Square     1     10.3226    0.0013
Mantel-Haenszel Chi-Square     1     11.7018    0.0006
Phi Coefficient                     -0.3849
Contingency Coefficient              0.3592
Cramer's V                          -0.3849

Sample Size = 80
```

### Output Explanation (Part B. 2x2 Table 2)

Since the p-value resulting from the SAS analysis (0.0006) is less than 0.05, one can reject the null hypothesis. Therefore, the locus of dependency in text editor satisfaction and computer experience found in the overall 3x2 contingency table in Example 7 can be determined. User satisfaction with the experimental text editor was significantly dependent upon the amount of computer experience when users classified as having high and medium experience were combined and compared to users classified as having low computer experience.

## Example 9: McNemar Change Test

**(Click in this red rectangle to see the Reference Notes on Example 9.)**

### Problem

Location in Williges (2006) Table of Contents
Section 2, Topic 5. Analysis of Nominal Data, Part 5.3.1. McNemar Change Test

Page(s) in Williges (2006) Reference Material: 198

Problem Description
50 people stated their preference for either Hearing Protector A or B before and after using each protector on the job for one week. To counterbalance order of use, half the people used Hearing Protector A on the job first and the other half used Hearing Protector B first. Did trial use of the hearing protectors change their preference ($p < 0.05$)?

Context/Purpose
Determine differences between before and after use preferences of two types of hearing protectors.

Statistical Decision Criteria
Since each subject used both hearing protectors, a within-subjects McNemar Change Test based on frequency data of a before and after scenario is appropriate. The 0.05 level of significance is used.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options pageno=1 nodate nocenter;
title 'Example 9: McNemar Change Test';
data protector;
input Before $ After $ count;
lines;
A A 13
A B 26
B B 5
B A 6
;
proc freq data=protector;
tables Before*After/agree alpha=0.05;
weight count;
quit;
```

### SAS Output**

***Note: SAS calculates the Pearson Chi-square statistic without using the more conservative Yate's Correction for Continuity. See the SAS Institute (2004) online documentation for the formula used to calculate the S statistic in the McNemar Change Test. The example in the Williges (2006) reference used*

*the Yate's correction. If the uncorrected formula were used, the Williges (2006) results would match the SAS program output.*

```
Example 9: McNemar Change Test

The FREQ Procedure

Table of Before by After

Before     After

Frequency
Percent
Row Pct
Col Pct   |A         |B         |  Total
----------+----------+----------+
A         |       13 |       26 |     39
          |    26.00 |    52.00 |  78.00
          |    33.33 |    66.67 |
          |    68.42 |    83.87 |
----------+----------+----------+
B         |        6 |        5 |     11
          |    12.00 |    10.00 |  22.00
          |    54.55 |    45.45 |
          |    31.58 |    16.13 |
----------+----------+----------+
Total           19         31        50
              38.00      62.00    100.00

Statistics for Table of Before by After

     McNemar's Test
-----------------------
Statistic (S)    12.5000
DF                     1
Pr > S            0.0004


    Simple Kappa Coefficient
--------------------------------
Kappa                  -0.1283
ASE                     0.1065
95% Lower Conf Limit   -0.3371
95% Upper Conf Limit    0.0804

Sample Size = 50
```

## Output Explanation

Since the p-value resulting from the SAS analysis (0.0004) is less than 0.05, one can reject the null hypothesis. Trial use of the hearing protectors did result in a change in preference before and after use. A significant number of the people changed their preference after using both hearing protectors.

## Example 10: Cochran Q Test

**(Click in this red rectangle to see the Reference Notes on Example 10.)**

### Problem

Location in Williges (2006) Table of Contents
Section 2, Topic 5. Analysis of Nominal Data, Part 5.3.2. Cochran Q Test

Page(s) in Williges (2006) Reference Material: 200 – 201

Problem Description
15 experienced photo interpreters viewed a series of photographs under three enhancement procedures and rated each procedure as “acceptable = 1” or “unacceptable = 0”. Are the three procedures rated equally? ($p < 0.001$)

Context/Purpose
Determine if there is a difference among the frequency of acceptability ratings given by the photo interpreters who evaluated each of three photo enhancement procedures.

Statistical Decision Criteria
Since frequency data among more than two related samples are being compared, a Cochran Q Test conducted at the 0.01 level of significance is appropriate to use with within-subjects nominal data.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 10: Cochran Q Test';
data photos;
input Proc1 Proc2 Proc3;
lines;
0 1 0
1 1 1
0 1 1
0 1 1
0 1 1
0 1 1
1 1 1
0 1 1
0 0 1
0 1 1
1 1 1
0 1 1
1 1 1
0 1 1
0 1 1
;
proc freq data=photos;
tables Proc1 Proc2 Proc3/nocum alpha=0.001;
```

```
tables Proc1*Proc2*Proc3/agree noprint alpha=0.001;
quit;
```

### SAS Output

```
Example 10: Cochran Q Test

The FREQ Procedure

Proc1    Frequency     Percent
──────────────────────────────
    0           11       73.33
    1            4       26.67


Proc2    Frequency     Percent
──────────────────────────────
    0            1        6.67
    1           14       93.33


Proc3    Frequency     Percent
──────────────────────────────
    0            1        6.67
    1           14       93.33


 Cochran's Q, for Proc1
   by Proc2 by Proc3
──────────────────────
Statistic (Q)    18.1818
DF                     2
Pr > Q            0.0001


Total Sample Size = 15
```

### Output Explanation

Since the p-value resulting from the SAS analysis (0.0001) is less than 0.001, one can reject the null hypothesis. Therefore, the frequency of acceptability ratings is significantly different among the three photo enhancement procedures. In order to determine these differences, one would need to perform a series of six McNemar Change Tests on all the paired-comparisons among the three procedures.

## Example 11: Kolmogorov-Smirnov Tests

**(Click in this red rectangle to see the Reference Notes on Example 11.)**

### Problem

Location in Williges (2006) Table of Contents
Section 2, Topic 6. Analysis of Ordinal Data, Part 6.2.1. Kolmogorov-Smirnov Tests

Page(s) in Williges (2006) Reference Material: 210 – 211

Problem Description
25 professional photographers and 30 nonprofessional photographers rated the acceptability of 25 photographs taken by an experimental camera on a 7 point Likert –type scale. Median acceptability ratings of 25 photographs were determined for each individual. Did the nonprofessionals give significantly higher median ratings of acceptability ($p < 0.01$)?

Context/Purpose
Determine if the median acceptability ratings given by non-professional photographers are higher than those of the professional photographers.

Statistical Decision Criteria
Use the Kolmogorov-Smirnov test because only two groups of between-subjects ordinal data are being compared at the 0.01 level of significance.

### Observed Data

The individual and median photo acceptability ratings are shown in the following tables for each professional and nonprofessional photographer, respectively. The frequency of median ratings at each of the 7 points on the Likert-type rating scale is used in the Kolmogorov-Smirnov Test.

| Professional | Acceptability of Photograph | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Photographer | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | Median |
| 1 | 1 | 3 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 1 | 1 |
| 2 | 1 | 5 | 3 | 7 | 4 | 3 | 3 | 2 | 3 | 5 | 6 | 3 | 1 | 3 | 5 | 7 | 3 | 2 | 1 | 2 | 6 | 3 | 4 | 3 | 2 | 3 |
| 3 | 5 | 1 | 6 | 5 | 4 | 6 | 7 | 2 | 7 | 5 | 6 | 3 | 7 | 5 | 7 | 6 | 2 | 6 | 4 | 7 | 6 | 4 | 5 | 6 | 7 | 6 |
| 4 | 4 | 4 | 2 | 2 | 1 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 3 | 2 | 1 | 1 | 1 | 1 | 3 | 4 | 1 | 1 |
| 5 | 3 | 4 | 6 | 2 | 4 | 5 | 5 | 3 | 4 | 1 | 7 | 6 | 3 | 4 | 5 | 6 | 2 | 6 | 1 | 7 | 5 | 4 | 6 | 5 | 6 | 5 |
| 6 | 3 | 2 | 4 | 1 | 6 | 2 | 3 | 1 | 1 | 5 | 2 | 3 | 1 | 2 | 6 | 2 | 3 | 1 | 4 | 2 | 3 | 7 | 4 | 1 | 2 | 2 |
| 7 | 5 | 3 | 4 | 1 | 6 | 3 | 1 | 5 | 4 | 2 | 3 | 1 | 5 | 4 | 2 | 7 | 3 | 4 | 5 | 7 | 6 | 2 | 3 | 1 | 4 | 4 |
| 8 | 2 | 3 | 4 | 1 | 5 | 1 | 2 | 4 | 3 | 1 | 2 | 4 | 6 | 5 | 2 | 3 | 1 | 2 | 2 | 7 | 5 | 6 | 2 | 1 | 1 | 2 |
| 9 | 3 | 2 | 2 | 5 | 1 | 3 | 2 | 6 | 2 | 3 | 3 | 1 | 2 | 4 | 5 | 1 | 2 | 3 | 1 | 7 | 4 | 3 | 1 | 2 | 1 | 2 |
| 10 | 1 | 2 | 1 | 1 | 1 | 3 | 5 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 1 |
| 11 | 5 | 4 | 6 | 1 | 6 | 7 | 5 | 7 | 6 | 4 | 1 | 6 | 7 | 5 | 6 | 2 | 4 | 7 | 6 | 5 | 7 | 2 | 3 | 6 | 7 | 6 |
| 12 | 1 | 5 | 3 | 6 | 4 | 3 | 7 | 6 | 7 | 5 | 4 | 5 | 1 | 2 | 4 | 5 | 3 | 7 | 2 | 5 | 6 | 5 | 4 | 1 | 5 | 5 |
| 13 | 3 | 6 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 3 | 2 | 1 | 1 | 1 | 2 | 3 | 4 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| 14 | 3 | 2 | 2 | 1 | 1 | 5 | 2 | 6 | 1 | 2 | 1 | 1 | 1 | 1 | 3 | 2 | 5 | 1 | 1 | 1 | 1 | 2 | 3 | 1 | 1 | 1 |
| 15 | 1 | 1 | 2 | 4 | 3 | 5 | 2 | 4 | 2 | 6 | 7 | 2 | 2 | 1 | 3 | 1 | 2 | 4 | 6 | 1 | 3 | 6 | 2 | 1 | 3 | 2 |
| 16 | 7 | 7 | 4 | 7 | 6 | 7 | 3 | 7 | 2 | 5 | 7 | 7 | 6 | 7 | 5 | 4 | 7 | 7 | 6 | 7 | 2 | 7 | 4 | 6 | 7 | 7 |
| 17 | 7 | 2 | 4 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 4 | 3 | 1 | 1 | 1 | 2 | 1 |
| 18 | 6 | 3 | 2 | 4 | 1 | 1 | 3 | 4 | 1 | 2 | 5 | 4 | 2 | 7 | 6 | 1 | 2 | 1 | 3 | 2 | 1 | 1 | 3 | 4 | 2 | 2 |
| 19 | 5 | 2 | 4 | 1 | 2 | 4 | 6 | 2 | 7 | 6 | 4 | 6 | 3 | 1 | 5 | 5 | 6 | 4 | 5 | 7 | 4 | 2 | 6 | 5 | 7 | 5 |
| 20 | 2 | 1 | 3 | 1 | 1 | 4 | 2 | 1 | 2 | 1 | 1 | 1 | 3 | 3 | 1 | 2 | 5 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 |
| 21 | 3 | 1 | 2 | 4 | 2 | 3 | 3 | 1 | 6 | 1 | 4 | 7 | 4 | 2 | 1 | 3 | 1 | 5 | 1 | 2 | 3 | 1 | 3 | 2 | 1 | 2 |
| 22 | 1 | 5 | 7 | 2 | 3 | 5 | 3 | 6 | 7 | 5 | 1 | 4 | 7 | 2 | 3 | 4 | 5 | 6 | 7 | 2 | 5 | 2 | 5 | 3 | 5 | 5 |
| 23 | 3 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 5 | 2 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 1 |
| 24 | 5 | 2 | 3 | 1 | 2 | 1 | 3 | 2 | 1 | 1 | 1 | 4 | 1 | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 1 |
| 25 | 4 | 3 | 5 | 2 | 6 | 2 | 5 | 3 | 1 | 3 | 5 | 1 | 4 | 7 | 6 | 2 | 4 | 5 | 7 | 2 | 3 | 4 | 3 | 2 | 5 | 4 |

### SAS Output**

***Note: The SAS output provides only D and tests the significance of the asymptotic KSa value. See the SAS Institute (2004) online documentation for a detailed description of these calculations. The SAS program does not calculate the Goodman Chi-Square statistic. To obtain the Goodman Chi-Square, use $D^2$ in the formula presented in the Williges (2006) reference.*

```
Example 11: Kolmogorov-Smirnov Test

The NPAR1WAY Procedure

   Kolmogorov-Smirnov Test for Variable rating
           Classified by Variable Group

                      EDF at    Deviation from Mean
Group        N       Maximum        at Maximum
_____________________________________________________
P           25      0.600000          1.272727
N           30      0.133333         -1.161836
Total       55      0.345455

   Maximum Deviation Occurred at Observation 9
         Value of rating at Maximum = 2.0

Kolmogorov-Smirnov Two-Sample Test (Asymptotic)
KS   0.232367    D         0.466667
KSa  1.723281    Pr > KSa  0.0053
```

### Output Explanation

The Asymptotic Kolmogorov-Smirnov Test (KSa) shown in the SAS output is significant at the 0.0053 level. The observed D statistic (0.46667) can be used to calculate the Goodman Chi-Square statistic. The resulting Goodman Chi-Square statistic (11.896) is larger than the tabled chi-square (9.21) at the 0.01 level of significance (Williges, 2005). So, both calculations yield significant results. One can reject the null hypothesis, which means that the non-professional photographers gave significantly higher ratings of acceptability than those of the professional photographers ($p < 0.01$).

## Example 12: Kruskal-Wallis One-Way ANOVA

**(Click in this red rectangle to see the Reference Notes on Example 12.)**

### Problem

Location in Williges (2006) Table of Contents
Section 2, Topic 6. Analysis of Ordinal Data, Part 6.2.2. Kruskal-Wallis One-Way ANOVA

Page(s) in Williges (2006) Reference Material: 215 – 216

Problem Description
A between-subjects design (n=6) was used to compare original learning by lecture, text, and multimedia instruction. Every trainee rated their overall satisfaction with the training on a 9-point scale. Did satisfaction differ across the three methods of training ($p < 0.05$)?

Context/Purpose
Determine if there is a difference in satisfaction across three types of multimedia instruction.

Statistical Decision Criteria
Use a Kruskal-Wallis One-Way ANOVA since comparisons are to be made among three or more independent samples of between-subjects ordinal data at the 0.05 level of significance.

### SAS Input**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

***Note: One must enter the rank order of all rating scores as the input into SAS program as shown in the Williges (2006) reference.*

```
options nodate nocenter pageno=1;
title 'Example 12: Kruskal-Wallis One-Way ANOVA';
data learning;
input Type $ Rank;
lines;
Lecture 13
Lecture 5.5
Lecture 7.5
Lecture 10
Lecture 13
Lecture 3.5
Multimedia 13
Multimedia 17
Multimedia 17
Multimedia 10
Multimedia 15
Multimedia 17
Text 3.5
Text 10
Text 7.5
Text 5.5
Text 2
```

```
Text 1
;
proc npar1way data=learning wilcoxon correct=no;
class Type;
var Rank;
exact wilcoxon /alpha=.05;
quit;
```

## SAS Output

```
Example 12: Kruskal-Wallis One-Way ANOVA

The NPAR1WAY Procedure

             Wilcoxon Scores (Rank Sums) for Variable Rank
                      Classified by Variable Type

                        Sum of      Expected       Std Dev          Mean
Type           N        Scores      Under H0      Under H0         Score
-------------------------------------------------------------------------
Lecture        6         52.50          57.0     10.594116      8.750000
Multimed       6         89.00          57.0     10.594116     14.833333
Text           6         29.50          57.0     10.594116      4.916667

                  Average scores were used for ties.


   Kruskal-Wallis Test

Chi-Square         10.6948
DF                       2
Pr > Chi-Square     0.0048


Monte Carlo Estimate for the Exact Test

Pr >= Chi-Square
Estimate                    0.0013
95% Lower Conf Limit     5.937E-04
95% Upper Conf Limit        0.0020

Number of Samples             10000
Initial Seed                  79643
```

## Output Explanation

Note that the SAS program automatically calculates the Kruskal-Wallis test based on tied ranks (KW = 10.6948) if ties exist in the data set. Since the p-value resulting from the SAS analysis (0.0048) is less than 0.05, one can reject the null hypothesis. Consequently, there is a significant difference in user satisfaction ratings among the three training methods. To determine the locus of these differences, additional post hoc Z tests as described in the Williges (2006) reference are necessary.

# Example 13: Wilcoxon Signed Ranks Test

## (Click in this red rectangle to see the Reference Notes on Example 13.)

## Problem

Location in Williges (2006) Table of Contents
Section 2, Topic 6. Analysis of Ordinal Data, Part 6.3.1. Wilcoxon Signed Ranks Test

Page(s) in Williges (2006) Reference Material: 220 - 221

Problem Description
Two electronic communication methods, video conferencing and instant messaging, were evaluated in a real-time battlefield information system on four 9-Point Likert-type scales in terms of ease of use, effectiveness, timeliness, and convenience by each soldier. Are the two forms of communication significantly different in terms of overall acceptability as measured by the sum of these four ratings ($p < 0.05$)?

Context/Purpose
Determine if there is a difference among acceptability ratings of communication methods.

Statistical Decision Criteria
Use the Wilcoxon Signed Rank Test, because the data were sampled from two ordinal, within-subjects samples at the 0.05 level of significance.

### Observed Data

These two tables show the soldier ratings of ease of use, effectiveness, timeliness, and convenience of the video conferencing and the instant messaging communication systems, respectively. The rank order of the difference between these two sums is used in the Wilcoxon Signed Rank Test as shown in Williges (2006).

| Video Conferencing | | | | | |
|---|---|---|---|---|---|
| **Soldier** | **Ease of Use** | **Effectiveness** | **Timeliness** | **Convenience** | **Sum** |
| **1** | 8 | 7 | 8 | 6 | 29 |
| **2** | 4 | 3 | 6 | 4 | 17 |
| **3** | 2 | 1 | 3 | 2 | 8 |
| **4** | 5 | 2 | 6 | 8 | 21 |
| **5** | 9 | 8 | 7 | 9 | 33 |
| **6** | 7 | 6 | 8 | 9 | 30 |
| **7** | 6 | 4 | 9 | 6 | 25 |
| **8** | 5 | 6 | 7 | 6 | 24 |
| **9** | 4 | 1 | 4 | 6 | 15 |
| **10** | 3 | 2 | 4 | 1 | 10 |
| **11** | 9 | 8 | 9 | 8 | 34 |

| Instant Messaging | | | | | |
|---|---|---|---|---|---|
| **Soldier** | **Ease of Use** | **Effectiveness** | **Timeliness** | **Convenience** | **Sum** |
| **1** | 7 | 8 | 5 | 6 | 26 |
| **2** | 4 | 5 | 1 | 1 | 11 |
| **3** | 3 | 5 | 3 | 1 | 12 |
| **4** | 2 | 2 | 2 | 2 | 8 |
| **5** | 2 | 1 | 1 | 1 | 5 |
| **6** | 5 | 6 | 4 | 4 | 19 |
| **7** | 5 | 3 | 5 | 7 | 20 |
| **8** | 4 | 2 | 2 | 2 | 10 |
| **9** | 4 | 3 | 6 | 6 | 19 |
| **10** | 1 | 2 | 4 | 5 | 12 |
| **11** | 6 | 4 | 3 | 5 | 18 |

**SAS Input****

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

*** Note: The data input into SAS are the rank order of the differences provided by Williges (2006) and not the actual differences of the sum of the four ratings shown in the video conferencing and instant messaging tables.*

```
options nodate nocenter pageno=1;
title 'Example 13: Wilcoxon Signed Rank Test';
data communication;
input subjects rank;
lines;
1 2
2 6
3 -3.5
4 8
5 11
6 7
7 5
8 9
9 -3.5
10 -1
11 10
;
proc univariate data=communication alpha=0.05;
var rank;
quit;
```

## SAS Output***

*** Note: *The S statistic presented below can be related to the T+ statistic demonstrated in the Williges (2006) reference material by using the following formula S = (T+) – [(n*(n +1))/4] as described in the SAS Institute (2004) online documentation. In addition, the SAS program uses an exact distribution when N ≤ 20 and uses a Student's t distribution when N > 20, which differs from the stated approach in Siegel and Castellan (1988) which is used in the Williges (2006) reference.*

```
Example 13: Wilcoxon Signed Rank Test

The UNIVARIATE Procedure
Variable:  rank

                          Moments

N                          11    Sum Weights                11
Mean                 4.54545455    Sum Observations           50
Std Deviation        5.27472533    Variance            27.8227273
Skewness             -0.5411792    Kurtosis            -1.1907993
Uncorrected SS            505.5    Corrected SS        278.227273
Coeff Variation      116.043957    Std Error Mean      1.59038953


              Basic Statistical Measures

    Location                     Variability

Mean      4.54545     Std Deviation            5.27473
Median    6.00000     Variance                27.82273
Mode     -3.50000     Range                   14.50000
                      Interquartile Range     10.00000


           Tests for Location: Mu0=0

Test           -Statistic-    -----p Value------

Student's t    t  2.858076    Pr > |t|    0.0170
Sign           M       2.5    Pr >= |M|   0.2266
Signed Rank    S        25    Pr >= |S|   0.0234
```

## Output Explanation

Since the p-value resulting from the SAS analysis (0.0234) is less than 0.05, one can reject the null hypothesis. Therefore, there is a statistically significant difference between the overall acceptability ratings of communication methods, such that the video conferencing communication system was rated higher than the instant messaging communication system in terms of the sum of the ease of use, effectiveness, timeliness, and convenience ratings.

## Example 14: Friedman Two-Way ANOVA

**(Click in this red rectangle to see the Reference Notes on Example 14.)**

### Problem

Location in Williges (2006) Table of Contents
Section 2. Topic 6. Analysis of Ordinal Data, Part 6.3.2. Friedman Two-Way ANOVA

Page(s) in Williges (2006) Reference Material: 225

Problem Description
Five subjects performed a benchmark task using a new CAD program. After completing the task, users rated their satisfaction using QUIS, and median ratings were calculated for each of the four parts of the scale, i.e., I. Screen, II. Terminology, III. Learning, and IV. Capability. Did median satisfaction differ across the parts ($p < 0.05$)?

Context/Purpose
Determine if there is a difference in median satisfaction with the CAD program across the four parts of the QUIS rating scale.

Statistical Decision/Criteria
Use the Friedman Two-Way ANOVA at the 0.05 level of significance since there are more than two categories of within-subjects ordinal data.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 14: Friedman Two-Way ANOVA';
data CAD;
input subject Part Rating;
lines;
1 1 2
2 1 4
3 1 1
4 1 0
5 1 5
1 2 6
2 2 8
3 2 9
4 2 5
5 2 7
1 3 7
2 3 9
3 3 6
4 3 8
5 3 4
1 4 3
2 4 3
3 4 2
```

```
4 4 1
5 4 6
;
proc freq data=CAD;
tables subject*Part*Rating/cmh2 scores=rank noprint alpha=0.05;
quit;
```

### SAS Output**

*** Note: The Friedman $F_r$ statistic is identical to the Row Mean Scores Differ Value of the Cochran-Mantel-Henszel (CMH) Statistic when based on rank order data as discussed in the SAS Institute (2004) online documentation. Consequently, the SAS output provides only the Row Mean Scores Differ rather than $F_r$ used in the Williges (2006) reference.*

```
Example 14: Friedman Two-Way ANOVA

The FREQ Procedure

Summary Statistics for Part by Rating
Controlling for subject

   Cochran-Mantel-Haenszel Statistics (Based on Rank Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
------------------------------------------------------------------
    1        Nonzero Correlation        1      0.3840    0.5355
    2        Row Mean Scores Differ     3      8.2800    0.0406


Total Sample Size = 20
```

### Output Explanation

Since the p-value from the SAS analysis (0.0406) is less than 0.05, one can reject the null hypothesis. Therefore, a statistically significant difference among median ratings of the four parts of the QUIS was detected. To determine which of the four parts of the QUIS scale are significantly different, one would need to conduct additional post hoc paired-comparison Z tests as described in the Williges (2006) reference material.

# Section 3. Basic Analysis of Variance (ANOVA) Designs

## Example 15: One-Factor, Between-Subjects ANOVA

**(Click in this red rectangle to see the Reference Notes on Example 15.)**

### Problem

Location in Williges (2006) Table of Contents
Section 3, Topic 10. Between-Subjects ANOVA Designs, Part 10.1.1. One-Factor Example

Page(s) in Williges (2006) Reference Material: 313 – 323

Problem Description
The effect of various aspects of information in military command and control situations was evaluated in terms of a commander's situation awareness. Situation awareness was measured for each of four different commanders who received information characterized as unreliable (u), ambiguous (a), or conflicting (c). Each commander received only one of the three types of information. Do these three aspects of information have a significant effect on commander's situation awareness ($p < 0.05$)?

Context/Purpose
Determine whether or not information characterized either as unreliable, ambiguous, or conflicting has a significant effect on the evaluation of a commander's situation awareness.

Statistical Decision Criteria
Perform a one-factor, between-subjects (each type of information is given to a different commander) ANOVA at the 0.05 level of significance.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 15: One-Way, Between-Subjects';
data information;
input subject $ characterization $ response;
lines;
1 u 42
2 u 41
3 u 37
4 u 40
5 a 43
6 a 49
7 a 52
8 a 48
9 c 32
10 c 40
11 c 41
12 c 39
;
proc glm;
class  subject characterization;
```

```
model response= characterization subject(characterization);
means characterization/alpha=.05;
test h=characterization e=subject(characterization);
run;
quit;
```

## SAS Output

```
Example 15: One-Way, Between-Subjects

The GLM Procedure

        Class Level Information

Class                 Levels    Values

subject                    4    1 2 3 4

characterization           3    a c u

Number of Observations Read          12
Number of Observations Used          12


Dependent Variable: response

                                     Sum of
Source                       DF         Squares     Mean Square    F Value    Pr > F

Model                        11     330.0000000      30.0000000        .       .

Error                         0       0.0000000        .

Corrected Total              11     330.0000000


R-Square     Coeff Var      Root MSE    response Mean

1.000000           .              .         42.00000


Source                       DF       Type I SS     Mean Square    F Value    Pr > F

characterization              2     224.0000000     112.0000000        .       .
subject(characteriz)          9     106.0000000      11.7777778        .       .


Source                       DF     Type III SS     Mean Square    F Value    Pr > F

characterization              2     224.0000000     112.0000000        .       .
subject(characteriz)          9     106.0000000      11.7777778        .       .
```

| Level of characterization | N | response Mean | response Std Dev |
|---|---|---|---|
| a | 4 | 48.0000000 | 3.74165739 |
| c | 4 | 38.0000000 | 4.08248290 |
| u | 4 | 40.0000000 | 2.16024690 |

Dependent Variable: response

Tests of Hypotheses Using the Type III MS for subject(characteriz) as an Error Term

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **characterization** | **2** | **224.0000000** | **112.0000000** | **9.51** | **0.0060** |

## Output Explanation

The obtained level of significance in SAS (i.e., p = 0.0060) is less than 0.05 level of significance, which leads to the rejection of the null hypothesis. This result indicates that there is a significant effect due to the three information characterizations on the commanders' spatial ability. Further analysis should be performed to determine which of the three types had effects on the commanders' situation awareness.

## Example 16: Two-Factor, Between-Subjects ANOVA

**(Click in this red rectangle to see the Reference Notes on Example 16.)**

### Problem

Location in Williges (2006) Table of Contents
Section 3, Topic 10. Between-Subjects ANOVA Designs, Part 10.2.4. Two-Factor Example

Page(s) in Williges (2006) Reference Material: 330 – 334

Problem Description
Readability of printed text on a computer screen was evaluated in terms of two fonts (Helvetica and Old English) and number of words displayed per line (10, 20, or 30 words per line). Four different subjects read one particular combination of these two factors and their reading comprehension was tested. Did either of these two factors or the interaction between them have a significant effect on reading comprehension ($p < 0.01$)?

Context/Purpose
Determine if the fonts, words displayed per line, or the interaction of these two factors have a significant effect on reading comprehension.

Statistical Decision Criteria
Conduct a two-factor (font and words per line), between-subjects (each subject is given a different treatment combination) ANOVA to determine if there are any significant differences at the 0.01 significance level.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 16: Two-Factor, Between-Subjects';
data six;
input Subject $ Font $ Words $ Response;
lines;
1 H 10 46
2 H 10 50
3 H 10 49
4 H 10 47
5 OE 10 47
6 OE 10 46
7 OE 10 50
8 OE 10 44
9 H 20 49
10 H 20 52
11 H 20 54
12 H 20 48
13 OE 20 39
14 OE 20 44
15 OE 20 38
```

```
16 OE 20 45
17 H 30 50
18 H 30 47
19 H 30 49
20 H 30 52
21 OE 30 35
22 OE 30 42
23 OE 30 39
24 OE 30 40
;
proc glm;
class Font Words Subject;
model Response = Font Words Subject(Font*Words) Font*Words;
means Font Words Font*Words/alpha=0.01;
test h=Font e=Subject(Font*Words);
test h=Words e=Subject(Font*Words);
test h=Font*Words e=Subject(Font*Words);
run;
quit;
```

## SAS Output

```
Example 16: Two-Factor, Between-Subjects

The GLM Procedure

     Class Level Information

Class         Levels    Values

Font               2    H OE

Words              3    10 20 30

Subject           24    1 10 11 12 13 14 15 16 17 18 19 2 20 21 22 23 24 3 4 5 6 7 8 9

Number of Observations Read          24
Number of Observations Used          24

Dependent Variable: Response

                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F

Model                       23     561.8333333      24.4275362        .       .

Error                        0       0.0000000        .

Corrected Total             23     561.8333333


R-Square     Coeff Var      Root MSE    Response Mean

1.000000           .              .         45.91667
```

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Font | 1 | 294.0000000 | 294.0000000 | . | . |
| Words | 2 | 39.5833333 | 19.7916667 | . | . |
| Subject(Font*Words) | 20 | 228.2500000 | 11.4125000 | . | . |
| Font*Words | 0 | 0.0000000 | . | . | . |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Font | 1 | 294.0000000 | 294.0000000 | . | . |
| Words | 2 | 39.5833333 | 19.7916667 | . | . |
| Subject(Font*Words) | 18 | 127.5000000 | 7.0833333 | . | . |
| Font*Words | 2 | 100.7500000 | 50.3750000 | . | . |

The GLM Procedure

| Level of Font | N | Response Mean | Response Std Dev |
|---|---|---|---|
| H | 12 | 49.4166667 | 2.35326981 |
| OE | 12 | 42.4166667 | 4.33711956 |

| Level of Words | N | Response Mean | Response Std Dev |
|---|---|---|---|
| 10 | 8 | 47.3750000 | 2.13390989 |
| 20 | 8 | 46.1250000 | 5.74300817 |
| 30 | 8 | 44.2500000 | 6.08863109 |

| Level of Font | Level of Words | N | Response Mean | Response Std Dev |
|---|---|---|---|---|
| H | 10 | 4 | 48.0000000 | 1.82574186 |
| H | 20 | 4 | 50.7500000 | 2.75378527 |
| H | 30 | 4 | 49.5000000 | 2.08166600 |
| OE | 10 | 4 | 46.7500000 | 2.50000000 |
| OE | 20 | 4 | 41.5000000 | 3.51188458 |
| OE | 30 | 4 | 39.0000000 | 2.94392029 |

Dependent Variable: Response

Tests of Hypotheses Using the Type III MS for Subject(Font*Words) as an Error Term

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Font** | **1** | **294.0000000** | **294.0000000** | **41.51** | **<.0001** |
| **Words** | **2** | **39.5833333** | **19.7916667** | **2.79** | **0.0877** |
| **Font*Words** | **2** | **100.7500000** | **50.3750000** | **7.11** | **0.0053** |

**Output Explanation**

The p-value of the font type (0.0001) is less than 0.01, leading to the rejection of the null hypothesis. The p-value of the number of words displayed per line (0.0877) is greater than 0.01, indicating that it does not have a significant effect on reading comprehension. The p-value of the interaction (0.0053) is also less than 0.01 which again results in rejection of the null hypothesis. Consequently, both the font type and the interaction have a significant effect on reading comprehension. To determine which type of font has a greater effect on reading comprehension, one would need to conduct post hoc analyses.

## Example 17: Planned Comparisons

**(Click in this red rectangle to see the Reference Notes on Example 17.)**

### Problem

Location in Williges (2006) Table of Contents
Section 3, Topic 11. Analysis of Comparisons and Interactions, Part 11.1.3.1 Planned Comparisons

Page(s) in Williges (2006) Reference Material: 353 – 359

Problem Description
The average number of seconds for 12 soldiers to locate a position on a standard black and white navigational map (1) was compared to 12 other soldiers using an experimental colored map (2), and 12 other soldiers using an experimental 3-D map (3). Four tests of significant difference in location time were planned: standard verses color, standard versus 3-D, color versus 3-D, and standard versus the average of color and 3-D maps. Which differences were significant ($p < 0.05$)?

Context/Purpose
Determine which of the four planned comparisons show a significant difference on the soldiers' ability to locate a position on a map. The first three are simple comparisons and the fourth is a complex comparison.

Statistical Decision Criteria
Use the contrast statement in GLM to test the four contrasts and use the Bonferroni t (Dunn) test to determine which of the three paired differences are significant at the 0.05 level.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 17: Planned Comparisons';
data location;
input subject $ map $ response;
lines;
1 1 5
2 1 3
3 1 4
4 1 3
5 1 7
6 1 6
7 1 7
8 1 2
9 1 5
10 1 4
11 1 6
12 1 2
1 2 6
```

```
2 2 4.3
3 2 3
4 2 5
5 2 3.8
6 2 5.2
7 2 4
8 2 5.5
9 2 3
10 2 4
11 2 4
12 2 5
1 3 5
2 3 7
3 3 9
4 3 4
5 3 6
6 3 8
7 3 11
8 3 6
9 3 5
10 3 9
11 3 10
12 3 4
;
proc glm;
class map;
model response = map;
lsmeans map/bon alpha=0.05;
contrast 'D1' map 1 -1 0;
contrast 'D2' map 1 0 -1;
contrast 'D3' map 0 1 -1;
contrast 'D4' map 2 -1 -1;
estimate 'CD1' map 1 -1 0;
estimate 'CD2' map 1 0 -1;
estimate 'CD3' map 0 1 -1;
run;
quit;
```

### SAS Output

```
Example 17: Planned Comparisons

The GLM Procedure

   Class Level Information

Class         Levels    Values

map                3    1 2 3


Number of Observations Read          36
Number of Observations Used          36
```

Dependent Variable: response

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 52.0800000 | 26.0400000 | 8.04 | 0.0014 |
| **Error** | **33** | **106.9000000** | **3.2393939** | | |
| **Corrected Total** | **35** | **158.9800000** | | | |

| R-Square | Coeff Var | Root MSE | response Mean |
|---|---|---|---|
| 0.327588 | 33.95909 | 1.799832 | 5.300000 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| map | 2 | 52.08000000 | 26.04000000 | 8.04 | 0.0014 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **map** | **2** | **52.08000000** | **26.04000000** | **8.04** | **0.0014** |

Least Squares Means

| map | response LSMEAN |
|---|---|
| 1 | 4.50000000 |
| 2 | 4.40000000 |
| 3 | 7.00000000 |

Bonferroni (Dunn) t Tests for response

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 33 |
| Error Mean Square | 3.239394 |
| **Critical Value of t** | **2.52221** |
| Minimum Significant Difference | 1.8533 |

Means with the same letter are not significantly different.

Bon
Grouping Mean N map

A 7.0000 12 3

B 4.5000 12 1

B 4.4000 12 2

Dependent Variable: response

| Contrast | DF | Contrast SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| D1 | 1 | 0.06000000 | 0.06000000 | 0.02 | 0.8926 |
| D2 | 1 | 37.50000000 | 37.50000000 | 11.58 | 0.0018 |
| D3 | 1 | 40.56000000 | 40.56000000 | 12.52 | 0.0012 |
| D4 | 1 | 11.52000000 | 11.52000000 | 3.56 | 0.0682 |

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| D1 | 0.10000000 | 0.73477819 | 0.14 | 0.8926 |
| D2 | -2.50000000 | 0.73477819 | -3.40 | 0.0018 |
| D3 | -2.60000000 | 0.73477819 | -3.54 | 0.0012 |

### Output Explanation

The planned F-test results in two significant comparisons by using the contrast statement. The comparison (D2) of standard versus 3-D displays is significant because the p-value (0.0018) is less than the specified significance (0.05). The comparison (D3) of color versus 3-D displays is significant since the p-value (0.0012) is less than the specified significance level (0.05). These comparisons have a significant effect on the location of the positions. The other two comparisons were not found to be significant because the p-values were greater than the significance value (0.05). The complex comparison (D4) is not significant at the 0.05 level since the p-value (0.0682) is larger. Note that SAS uses the estimate function to determine the critical differences of the means as opposed to using treatment totals. The Bonferroni t-test results in significant differences between the standard and 3-D display (D2) and between the color and 3-D display (D3). This result is consistent with the results of the planned F-test and critical differences. The results of this analysis indicate that there is a significant difference in the location of positions when the soldiers used the 3-D display as opposed to the standard and color displays.

## Example 18: Unplanned Comparisons

**(Click in this red rectangle to see the Reference Notes on Example 18.)**

### Problem

Location in Williges (2006) Table of Contents
Section 3, Topic 11. Analysis of Comparisons and Interactions, Part 11.1.4. Unplanned Comparisons

Page(s) in Williges (2006) Reference Material: 362 – 373

Problem Description
Proprioceptive, visual, sound, and voice modes of presenting information were evaluated by 24 soldiers. One of these four modes of information was randomly assigned to 6 soldiers using wearable computers during training maneuvers. There was an overall significant mode difference in minutes to complete the training maneuver ($p < 0.05$). Which communication modes were significantly different from each other?

Context/Purpose
Determine which of the four modes of communication were different from each other by conducting a series of post hoc paired comparisons to isolate the significant main effect of information presentation mode.

Statistical Decision Criteria
Perform a one-way, between-subjects ANOVA on the four modes of communication with additional tests of Least Significant Difference, Bonferroni, Scheffe, Tukey, Dunnett, and Student Newman-Keuls at the 0.05 level of significance. The visual mode is the control for the Dunnett test.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 18: Unplanned Comparisons';
data wearable;
input mode $ response;
lines;
proprio 10
proprio 13
proprio 14
proprio 14
proprio 15
proprio 17
visual 10
visual 10
visual 13
visual 14
visual 12
visual 15
```

```
voice 15
voice 15
voice 16
voice 17
voice 16
voice 18
sound 13
sound 14
sound 12
sound 18
sound 16
sound 19
;
proc glm;
class mode;
model response = mode;
means mode/lsd bon scheffe tukey dunnett('visual') snk alpha=0.05;
run;
quit;
```

### SAS Output**

***Note: The output displayed below has been re-ordered from the original SAS output to make the results easier to read.*

```
Example 18: Unplanned Comparisons

The GLM Procedure

              Class Level Information

Class         Levels    Values

mode               4    proprio sound visual voice


Number of Observations Read          24
Number of Observations Used          24
```

The GLM Procedure

Dependent Variable: response

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | **3** | **51.5000000** | **17.1666667** | **3.64** | **0.0304** |
| **Error** | **20** | **94.3333333** | **4.7166667** | | |
| **Corrected Total** | **23** | **145.8333333** | | | |

| R-Square | Coeff Var | Root MSE | response Mean |
|---|---|---|---|
| 0.353143 | 15.06443 | 2.171789 | 14.41667 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| mode | 3 | 51.50000000 | 17.16666667 | 3.64 | 0.0304 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **mode** | **3** | **51.50000000** | **17.16666667** | **3.64** | **0.0304** |

t Tests (LSD) for response

NOTE: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 20 |
| Error Mean Square | 4.716667 |
| Critical Value of t | 2.08596 |
| **Least Significant Difference** | **2.6156** |

Comparisons significant at the 0.05 level are indicated by ***.

| mode Comparison | Difference Between Means | 95% Confidence Limits | | |
|---|---|---|---|---|
| voice - sound | 0.833 | -1.782 | 3.449 | |
| voice - proprio | 2.333 | -0.282 | 4.949 | |
| **voice - visual** | **3.833** | **1.218** | **6.449** | *** |
| sound - voice | -0.833 | -3.449 | 1.782 | |
| sound - proprio | 1.500 | -1.116 | 4.116 | |
| **sound - visual** | **3.000** | **0.384** | **5.616** | *** |
| proprio - voice | -2.333 | -4.949 | 0.282 | |
| proprio - sound | -1.500 | -4.116 | 1.116 | |
| proprio - visual | 1.500 | -1.116 | 4.116 | |
| visual - voice | -3.833 | -6.449 | -1.218 | *** |
| visual - sound | -3.000 | -5.616 | -0.384 | *** |
| visual - proprio | -1.500 | -4.116 | 1.116 | |

t Tests (LSD) for response

NOTE: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 20 |
| Error Mean Square | 4.716667 |
| Critical Value of t | 2.08596 |
| **Least Significant Difference** | **2.6156** |

Means with the same letter are not significantly different.

| t Grouping | Mean | N | mode |
|---|---|---|---|
| **A** | **16.167** | **6** | **voice** |
| **A** | **15.333** | **6** | **sound** |
| B A | 13.833 | 6 | proprio |
| **B** | **12.333** | **6** | **visual** |

Tukey's Studentized Range (HSD) Test for response

NOTE: This test controls the Type I experimentwise error rate.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 20 |
| Error Mean Square | 4.716667 |
| Critical Value of Studentized Range | 3.95829 |
| **Minimum Significant Difference** | **3.5095** |

Comparisons significant at the 0.05 level are indicated by ***.

| mode Comparison | Difference Between Means | Simultaneous 95% Confidence Limits | | |
|---|---|---|---|---|
| voice - sound | 0.833 | -2.676 | 4.343 | |
| voice - proprio | 2.333 | -1.176 | 5.843 | |
| **voice - visual** | **3.833** | **0.324** | **7.343** | *** |
| sound - voice | -0.833 | -4.343 | 2.676 | |
| sound - proprio | 1.500 | -2.010 | 5.010 | |
| sound - visual | 3.000 | -0.510 | 6.510 | |
| proprio - voice | -2.333 | -5.843 | 1.176 | |
| proprio - sound | -1.500 | -5.010 | 2.010 | |
| proprio - visual | 1.500 | -2.010 | 5.010 | |
| visual - voice | -3.833 | -7.343 | -0.324 | *** |
| visual - sound | -3.000 | -6.510 | 0.510 | |
| visual - proprio | -1.500 | -5.010 | 2.010 | |

Tukey's Studentized Range (HSD) Test for response

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 20 |
| Error Mean Square | 4.716667 |
| Critical Value of Studentized Range | 3.95829 |
| **Minimum Significant Difference** | **3.5095** |

Means with the same letter are not significantly different.

| Tukey Grouping | Mean | N | mode |
|---|---|---|---|
| **A** | **16.167** | **6** | **voice** |
| B A | 15.333 | 6 | sound |
| B A | 13.833 | 6 | proprio |
| **B** | **12.333** | **6** | **visual** |

Bonferroni (Dunn) t Tests for response

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than Tukey's for all pairwise comparisons.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 20 |
| Error Mean Square | 4.716667 |
| Critical Value of t | 2.92712 |
| **Minimum Significant Difference** | **3.6703** |

Comparisons significant at the 0.05 level are indicated by ***.

| mode Comparison | Difference Between Means | Simultaneous 95% Confidence Limits | | |
|---|---|---|---|---|
| voice - sound | 0.833 | -2.837 | 4.504 | |
| voice - proprio | 2.333 | -1.337 | 6.004 | |
| **voice - visual** | **3.833** | **0.163** | **7.504** | *** |
| sound - voice | -0.833 | -4.504 | 2.837 | |
| sound - proprio | 1.500 | -2.170 | 5.170 | |
| sound - visual | 3.000 | -0.670 | 6.670 | |
| proprio - voice | -2.333 | -6.004 | 1.337 | |
| proprio - sound | -1.500 | -5.170 | 2.170 | |
| proprio - visual | 1.500 | -2.170 | 5.170 | |
| visual - voice | -3.833 | -7.504 | -0.163 | *** |
| visual - sound | -3.000 | -6.670 | 0.670 | |
| visual - proprio | -1.500 | -5.170 | 2.170 | |

Bonferroni (Dunn) t Tests for response

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 20 |
| Error Mean Square | 4.716667 |
| Critical Value of t | 2.92712 |
| **Minimum Significant Difference** | **3.6703** |

Means with the same letter are not significantly different.

| Bon Grouping | | Mean | N | mode |
|---|---|---|---|---|
| | **A** | **16.167** | **6** | **voice** |
| B | A | 15.333 | 6 | sound |
| B | A | 13.833 | 6 | proprio |
| **B** | | **12.333** | **6** | **visual** |

Scheffe's Test for response

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than Tukey's for all pairwise comparisons.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 20 |
| Error Mean Square | 4.716667 |
| Critical Value of F | 3.09839 |
| **Minimum Significant Difference** | **3.8228** |

Comparisons significant at the 0.05 level are indicated by ***.

| mode Comparison | Difference Between Means | Simultaneous 95% Confidence Limits | | |
|---|---|---|---|---|
| voice - sound | 0.833 | -2.990 | 4.656 | |
| voice - proprio | 2.333 | -1.490 | 6.156 | |
| **voice - visual** | **3.833** | **0.010** | **7.656** | *** |
| sound - voice | -0.833 | -4.656 | 2.990 | |
| sound - proprio | 1.500 | -2.323 | 5.323 | |
| sound - visual | 3.000 | -0.823 | 6.823 | |
| proprio - voice | -2.333 | -6.156 | 1.490 | |
| proprio - sound | -1.500 | -5.323 | 2.323 | |
| proprio - visual | 1.500 | -2.323 | 5.323 | |
| visual - voice | -3.833 | -7.656 | -0.010 | *** |
| visual - sound | -3.000 | -6.823 | 0.823 | |
| visual - proprio | -1.500 | -5.323 | 2.323 | |

Scheffe's Test for response

NOTE: This test controls the Type I experimentwise error rate.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 20 |
| Error Mean Square | 4.716667 |
| Critical Value of F | 3.09839 |
| **Minimum Significant Difference** | **3.8228** |

Means with the same letter are not significantly different.

| Scheffe Grouping | Mean | N | mode |
|---|---|---|---|
| **A** | **16.167** | **6** | **voice** |
| B A | 15.333 | 6 | sound |
| B A | 13.833 | 6 | proprio |
| **B** | **12.333** | **6** | **visual** |

Dunnett's t Tests for response

NOTE: This test controls the Type I experimentwise error for comparisons of all treatments against
a control.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 20 |
| Error Mean Square | 4.716667 |
| Critical Value of Dunnett's t | 2.54043 |
| **Minimum Significant Difference** | **3.1854** |

Comparisons significant at the 0.05 level are indicated by ***.

| mode Comparison | Difference Between Means | Simultaneous 95% Confidence Limits | | |
|---|---|---|---|---|
| **voice - visual** | **3.833** | **0.648** | **7.019** | ******* |
| sound - visual | 3.000 | -0.185 | 6.185 | |
| proprio - visual | 1.500 | -1.685 | 4.685 | |

***Note: Since the Dunnett test compares only to the control (visual), only three output comparisons are given.*

Student-Newman-Keuls Test for response

NOTE: This test controls the Type I experimentwise error rate under the complete null hypothesis but not under partial null hypotheses.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 20 |
| Error Mean Square | 4.716667 |

| **Number of Means** | **2** | **3** | **4** |
|---|---|---|---|
| **Critical Range** | **2.6155553** | **3.1723012** | **3.5095392** |

Means with the same letter are not significantly different.

| SNK Grouping | Mean | N | mode |
|---|---|---|---|
| **A** | **16.167** | **6** | **voice** |
| B A | 15.333 | 6 | sound |
| B A | 13.833 | 6 | proprio |
| **B** | **12.333** | **6** | **visual** |

**Output Explanation**

The ANOVA output indicates that the mode of presentation does have a significant effect on time during the maneuver ($p = 0.03$). To determine which of these modes are significantly different from the others, post-hoc analyses were performed. The Least Significant Difference (LSD) test resulted in critical difference of 2.62 and significant differences between the voice and visual modes of communication as well as the sound and visual modes. The Tukey HSD test resulted in a significant difference between the voice and visual modes with a critical difference of 3.51. The Bonferroni test resulted in a critical difference of 3.67 and a significant difference between the voice and visual modes. The Scheffé test resulted in a critical difference of 3.82 and a significant difference between the voice and visual modes. The Dunnett test resulted in a critical difference of 3.19 and a significant difference between the voice and visual communication modes. The Student Newman-Keuls (SNK) test resulted in critical differences of 2.62, 3.17, and 3.51. The SNK also reports a significant difference between the voice and visual modes of communication. All of the post-hoc comparisons report that there is a significant difference between the voice and visual communication modes. However, the LSD test reports a second significant difference pair. This may have occurred because the LSD test is the most lax of those performed and therefore found more significant differences (Williges 2005). Since all of the tests resulted in a significant difference between the voice and visual modes of communication, it can be said that these two modes have a significant effect on time to complete the maneuver.

## Example 19: Analysis of Interactions

**(Click in this red rectangle to see the Reference Notes on Example 19.)**

### Problem

Location in Williges (2006) Table of Contents
Section 3, Topic 11. Analysis of Comparisons and Interactions, Part 11.2.1. Two-Factor Interaction Example Problem

Page(s) in Williges (2006) Reference Material: 377 – 397

Problem Description
Distributed and co-located teams evaluated four zoom percentages (0, 50, 100, 150%) of computer displays. An overall ANOVA resulted in a significant interaction ($p < 0.05$) between type of team and percent zoom in terms of the percentage of threat evaluations made correctly. Based on the mean values in this between-subjects design, where is the locus of the interaction in terms of improving team communication and collaboration?

Context/Purpose
Isolate and interpret the significant interaction between the type of team and the percent zoom of computer displays resulting from the overall ANOVA.

Statistical Decision Criteria
Conduct simple effects tests, trend analyses, and post hoc paired comparisons at the 0.01 level of significance to isolate the significant interaction effects.

### SAS Input (Part A. Simple Effects)

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 19A: Simple Effects of the Interaction';
data interactions;
input subject $ teams $ zoom $ response;
lines;
1 d 0 79
2 d 0 75
3 d 0 77
4 c 0 91
5 c 0 90
6 c 0 98
7 d 50 82
8 d 50 83
9 d 50 79
10 c 50 92
11 c 50 95
12 c 50 95
13 d 100 90
14 d 100 82
15 d 100 80
```

```
16 c 100 88
17 c 100 95
18 c 100 93
19 d 150 95
20 d 150 89
21 d 150 92
22 c 150 90
23 c 150 87
24 c 150 96
;
proc glm;
class teams zoom subject;
model response = teams zoom teams*zoom;
means teams zoom teams*zoom/alpha=0.05;
lsmeans teams*zoom/ slice=teams;
run;
quit;
```

### SAS Output (Part A. Simple Effects)

```
Example 19A: Simple Effects of the Interaction
1

The GLM Procedure

                              Class Level Information

Class         Levels    Values

teams              2    c d

zoom               4    0 100 150 50

subject           24    1 10 11 12 13 14 15 16 17 18 19 2 20 21 22 23 24 3 4 5 6 7 8 9


Number of Observations Read          24
Number of Observations Used          24
```

Dependent Variable: response

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 7 | 850.291667 | 121.470238 | 9.59 | 0.0001 |
| **Error** | **16** | **202.666667** | **12.666667** | | |
| **Corrected Total** | **23** | **1052.958333** | | | |

| R-Square | Coeff Var | Root MSE | response Mean |
|---|---|---|---|
| 0.807526 | 4.042434 | 3.559026 | 88.04167 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| teams | 1 | 477.0416667 | 477.0416667 | 37.66 | <.0001 |
| zoom | 3 | 128.1250000 | 42.7083333 | 3.37 | 0.0446 |
| teams*zoom | 3 | 245.1250000 | 81.7083333 | 6.45 | 0.0045 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **teams** | **1** | **477.0416667** | **477.0416667** | **37.66** | **<.0001** |
| **zoom** | **3** | **128.1250000** | **42.7083333** | **3.37** | **0.0446** |
| **teams*zoom** | **3** | **245.1250000** | **81.7083333** | **6.45** | **0.0045** |

| Level of teams | N | response Mean | response Std Dev |
|---|---|---|---|
| c | 12 | 92.5000000 | 3.39786029 |
| d | 12 | 83.5833333 | 6.38831794 |

| Level of zoom | N | response Mean | response Std Dev |
|---|---|---|---|
| 0 | 6 | 85.0000000 | 9.27361850 |
| 100 | 6 | 88.0000000 | 5.96657356 |
| 150 | 6 | 91.5000000 | 3.50713558 |
| 50 | 6 | 87.6666667 | 7.14609450 |

| Level of teams | Level of zoom | N | response Mean | response Std Dev |
|---|---|---|---|---|
| c | 0 | 3 | 93.0000000 | 4.35889894 |
| c | 100 | 3 | 92.0000000 | 3.60555128 |
| c | 150 | 3 | 91.0000000 | 4.58257569 |
| c | 50 | 3 | 94.0000000 | 1.73205081 |
| d | 0 | 3 | 77.0000000 | 2.00000000 |
| d | 100 | 3 | 84.0000000 | 5.29150262 |
| d | 150 | 3 | 92.0000000 | 3.00000000 |
| d | 50 | 3 | 81.3333333 | 2.08166600 |

Least Squares Means

| teams | zoom | response LSMEAN |
|---|---|---|
| c | 0 | 93.0000000 |
| c | 100 | 92.0000000 |
| c | 150 | 91.0000000 |
| c | 50 | 94.0000000 |
| d | 0 | 77.0000000 |
| d | 100 | 84.0000000 |
| d | 150 | 92.0000000 |
| d | 50 | 81.3333333 |

Least Squares Means

teams*zoom Effect Sliced by teams for response

| teams | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **c** | **3** | **15.000000** | **5.000000** | **0.39** | **0.7585** |
| **d** | **3** | **358.250000** | **119.416667** | **9.43** | **0.0008** |

### Output Explanation (Part A. Simple Effects)

The p-value (0.0008) for threat evaluation performance for distributed teams across various computer display zoom levels is less than 0.05, which leads to the rejection of the null hypothesis. There is not a significant effect display zoom percentage on threat evaluations of co-located teams, because the p-value (0.7585) is larger than 0.05. Consequently, the interaction is due to the effect of changes in the zoom level of computer displays used by distributed teams, not co-located teams.

**SAS Input (Part B. Trend Analysis)****

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

*** Note: The teams*zoom interaction contrasts uses the design model to determine the coefficients for the contrasts statements.*

```
options nodate nocenter pageno=1;
title 'Example 19B: Trend Analysis of the Interaction';
data interactions;
input teams $ zoom response;
lines;
d 0 79
d 0 75
d 0 77
c 0 91
c 0 90
c 0 98
d 50 82
d 50 83
d 50 79
c 50 92
c 50 95
c 50 95
d 100 90
d 100 82
d 100 80
c 100 88
c 100 95
c 100 93
d 150 95
d 150 89
d 150 92
c 150 90
c 150 87
c 150 96
;
proc glm;
class teams zoom;
model response = teams zoom teams*zoom;
lsmeans teams*zoom/alpha=0.05;
contrast 'Linear at teams=c' zoom -3 -1 1 3 teams*zoom -3 -1 1 3 0 0 0 0;
contrast 'Quadratic at teams=c' zoom 1 -1 -1 1 teams*zoom 1 -1 -1 1 0 0 0 0;
contrast 'Cubic at teams=c' zoom -1 3 -3 1 teams*zoom -1 3 -3 1 0 0 0 0;
contrast 'Linear at teams=d' zoom -3 -1 1 3 teams*zoom 0 0 0 0 -3 -1 1 3;
contrast 'Quadratic at teams=d' zoom 1 -1 -1 1 teams*zoom 0 0 0 0 1 -1 -1 1;
contrast 'Cubic at teams=d' zoom -1 3 -3 1 teams*zoom 0 0 0 0 -1 3 -3 1;
run;
quit;
```

## SAS Output (Part B. Trend Analysis)

```
Example 19B: Trend Analysis of the Interaction                              1

The GLM Procedure

       Class Level Information

Class         Levels    Values

teams              2    c d

zoom               4    0 50 100 150


Number of Observations Read          24
Number of Observations Used          24


Dependent Variable: response

                                   Sum of
Source                       DF         Squares     Mean Square    F Value    Pr > F

Model                         7      850.291667      121.470238       9.59    0.0001

Error                        16      202.666667       12.666667

Corrected Total              23     1052.958333


R-Square     Coeff Var      Root MSE    response Mean

0.807526      4.042434      3.559026         88.04167


Source                       DF       Type I SS     Mean Square    F Value    Pr > F

teams                         1     477.0416667     477.0416667      37.66    <.0001
zoom                          3     128.1250000      42.7083333       3.37    0.0446
teams*zoom                    3     245.1250000      81.7083333       6.45    0.0045


Source                       DF     Type III SS     Mean Square    F Value    Pr > F

teams                         1     477.0416667     477.0416667      37.66    <.0001
zoom                          3     128.1250000      42.7083333       3.37    0.0446
teams*zoom                    3     245.1250000      81.7083333       6.45    0.0045
```

Least Squares Means

| teams | zoom | response LSMEAN |
|---|---|---|
| c | 0 | 93.0000000 |
| c | 50 | 94.0000000 |
| c | 100 | 92.0000000 |
| c | 150 | 91.0000000 |
| d | 0 | 77.0000000 |
| d | 50 | 81.3333333 |
| d | 100 | 84.0000000 |
| d | 150 | 92.0000000 |

Dependent Variable: response

| Contrast | DF | Contrast SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Linear at teams=c** | **1** | **9.6000000** | **9.6000000** | **0.76** | **0.3969** |
| **Quadratic at teams=c** | **1** | **3.0000000** | **3.0000000** | **0.24** | **0.6331** |
| **Cubic at teams=c** | **1** | **2.4000000** | **2.4000000** | **0.19** | **0.6692** |
| **Linear at teams=d** | **1** | **340.8166667** | **340.8166667** | **26.91** | **<.0001** |
| **Quadratic at teams=d** | **1** | **10.0833333** | **10.0833333** | **0.80** | **0.3855** |
| **Cubic at teams=d** | **1** | **7.3500000** | **7.3500000** | **0.58** | **0.4573** |

### Output Explanation (Part B. Trend Analysis)

Only the p-value for the linear trend of distributed teams (<0.0001) is less than 0.05. Consequently, the significant interaction effect is due to a linear decrease in threat evaluations made by distributed teams as the percent of zoom level in computerized information displays increases.

**SAS Input (Part C. Newman-Keuls Paired Comparisons)****

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

*** Note: The Newman-Keuls analysis in SAS requires a reordering of the input data. The combinations here match those listed in increasing rank order as treatments on page 391 of the Williges (2006) reference.*

```
options nodate nocenter pageno=1;
title 'Example 19C: SNK Paired Comparisons of the Interaction;
data interactions;
input combination $ response;
lines;
1 79
1 75
1 77
2 82
2 83
2 79
3 90
3 82
3 80
4 90
4 87
4 96
5 95
5 89
5 92
6 88
6 95
6 93
7 91
7 90
7 98
8 92
8 95
8 95
;
proc glm;
class combination;
model response = combination;
means combination/snk alpha=.05;
run;
quit;
```

## SAS Output (Part C. Newman-Keuls Paired Comparisons)

```
Example 19C: SNK Paired Comparisons of the Interaction

          Class Level Information

Class            Levels    Values

combination           8    1 2 3 4 5 6 7 8


Number of Observations Read          24
Number of Observations Used          24


Dependent Variable: response

                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F

Model                        7      850.291667      121.470238       9.59    0.0001

Error                       16      202.666667       12.666667

Corrected Total             23     1052.958333


R-Square     Coeff Var      Root MSE    response Mean

0.807526      4.042434      3.559026         88.04167


Source                      DF       Type I SS     Mean Square    F Value    Pr > F

combination                  7     850.2916667     121.4702381       9.59    0.0001


Source                      DF     Type III SS     Mean Square    F Value    Pr > F

combination                  7     850.2916667     121.4702381       9.59    0.0001
```

```
Student-Newman-Keuls Test for response

NOTE: This test controls the Type I experimentwise error rate under the complete null hypothesis
but not under partial null hypotheses.


Alpha                          0.05
Error Degrees of Freedom         16
Error Mean Square          12.66667

Number of Means           2         3         4         5         6         7         8
Critical Range    6.1603024 7.4982682 8.3139319 8.9028283 9.3633532 9.7410311 10.060777


Means with the same letter are not significantly different.


 SNK
Groupi
  ng            Mean      N    combination

     A        94.000      3    8

     A        93.000      3    7

B    A        92.000      3    5

B    A        92.000      3    6

B    A        91.000      3    4

B    C        84.000      3    3

     C        81.333      3    2

     C        77.000      3    1
```

### Output Explanation (Part C. Newman-Keuls Paired Comparisons)

The Newman-Keuls analysis resulted in twelve significant comparisons including the (8,1), (8,2), (8,3), (7,1), (7,2), (7,3), (6,1), (6,2), (5,1), (5,2), (4,1), and (4,2) differences. Only three of these comparisons are unconfounded and have an effect on the interaction. Namely, the significant differences between distributed teams using 50% zoom and co-located teams using 50% zoom (8,2), the difference between distributed teams using 0% zoom and co-located teams using 0% zoom (7,1), and the difference between distributed teams using 150% zoom and distributed teams using 0% zoom (5,1). Consequently, computer display zoom only affects distributed teams who detect significantly fewer threats than co-located teams when teams only have access to 50% and 0% zoom displays. Note that the critical difference values calculated by SAS are the mean values not totals. To get the total values shown in the Williges (2006) reference, you would need to multiply these values by 2.33.

**SAS Input (Part D. LSD Test of Paired Comparisons)****

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

*****Note*: The combination designation corresponds to the AB combination of the original problem. For example combination 11 corresponds to A1, B1 which is Distributed 0%. This example must be coded into SAS as a one-way ANOVA to achieve the correct results for the LSD.*

```
options nodate nocenter pageno=1;
title 'Example 19D: LSD Paired Comparisons of the Interaction';
data interactions;
input ABcombination $ response;
lines;
11 79
11 75
11 77
21 91
21 90
21 98
12 82
12 83
12 79
22 92
22 95
22 95
13 90
13 82
13 80
23 88
23 95
23 93
14 95
14 89
14 92
24 90
24 87
24 96
;
proc glm;
class ABcombination;
model response = ABcombination;
means ABcombination/lsd alpha=0.05;
run;
quit;
```

## SAS Output (Part D. LSD Test of Paired Comparisons)

```
Example 19D: LSD Paired Comparisons of the Interaction
1

The GLM Procedure

              Class Level Information

Class            Levels    Values

ABcombination           8    11 12 13 14 21 22 23 24


Number of Observations Read          24
Number of Observations Used          24

Dependent Variable: response

                                     Sum of
Source                       DF         Squares     Mean Square    F Value    Pr > F

Model                         7      850.291667      121.470238       9.59    0.0001

Error                        16      202.666667       12.666667

Corrected Total              23     1052.958333


R-Square     Coeff Var      Root MSE    response Mean

0.807526      4.042434      3.559026         88.04167


Source                       DF       Type I SS     Mean Square    F Value    Pr > F

ABcombination                   7     850.2916667     121.4702381       9.59    0.0001


Source                       DF     Type III SS     Mean Square    F Value    Pr > F

ABcombination                   7     850.2916667     121.4702381       9.59    0.0001
```

```
t Tests (LSD) for response

NOTE: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.


Alpha                           0.05
Error Degrees of Freedom          16
Error Mean Square            12.66667
Critical Value of t           2.11991
Least Significant Difference   6.1603


Means with the same letter are not significantly different.


  t
Groupi
  ng            Mean      N    ABcombination

     A        94.000      3    22

     A        93.000      3    21

     A        92.000      3    23

     A        92.000      3    14

     A        91.000      3    24

     B        84.000      3    13

C    B        81.333      3    12

C             77.000      3    11
```

**Output Explanation (Part D. LSD Test of Paired Comparisons)**

The following pairs are significant because they do not have the same letter as indicated by SAS: (22,13) (22,12) (22,11) (21,13) (21,12) (21,11) (23,13) (23,12) (23,11) (14,13) (14,12) (14,11) (24,13) (24,12) (24,11) and (13,11). However, only the pairs (22,12) (21,11) (23,13) (14,13) (14,12) (14,11) and (13,11) are unconfounded pairs and contribute to the interaction. The significant differences in threat evaluation performance between unconfounded pairs in the LSD analysis are the same as those found in the adjusted Bonferroni t analysis. Note that the least significant difference value calculated by SAS is off by a factor of three. SAS calculates this value using means, while the calculations in the Williges (2006) reference were done using totals.

**SAS Input (Part E. Bonferroni t Paired Comparisons)****

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

*****Note*: SAS calculates the Bonferroni correction for all comparisons. It does not differentiate between confounded or unconfounded comparisons of the interaction.*

```
options nodate nocenter pageno=1;
title 'Example 19E: Bonferroni t Paired Comparisons of the Interaction';
data interactions;
input subject $ teams $ zoom $ response;
lines;
1 d 150 95
2 d 150 89
3 d 150 92
4 c 150 90
5 c 150 87
6 c 150 96
7 d 100 90
8 d 100 82
9 d 100 80
10 c 100 88
11 c 100 95
12 c 100 93
13 d 50 82
14 d 50 83
15 d 50 79
16 c 50 92
17 c 50 95
18 c 50 95
19 d 0 79
20 d 0 75
21 d 0 77
22 c 0 91
23 c 0 90
24 c 0 98
;
proc glm;
class teams zoom subject;
model response = teams zoom teams*zoom;
lsmeans teams*zoom/ pdiff adjust=bon alpha=0.05;
run;
quit;
```

## SAS Output (Part E. Bonferroni t Paired Comparisons)

```
Example 19E: Bonferroni t Paired Comparisons of the Interaction
1

The GLM Procedure

                              Class Level Information

Class         Levels    Values

teams              2    c d

zoom               4    0 100 150 50

subject           24    1 10 11 12 13 14 15 16 17 18 19 2 20 21 22 23 24 3 4 5 6 7 8 9


Number of Observations Read          24
Number of Observations Used          24


Dependent Variable: response

                                   Sum of
Source                     DF         Squares     Mean Square    F Value    Pr > F

Model                       7      850.291667      121.470238       9.59    0.0001

Error                      16      202.666667       12.666667

Corrected Total            23     1052.958333


R-Square     Coeff Var      Root MSE    response Mean

0.807526      4.042434      3.559026         88.04167


Source                     DF       Type I SS     Mean Square    F Value    Pr > F

teams                       1     477.0416667     477.0416667      37.66    <.0001
zoom                        3     128.1250000      42.7083333       3.37    0.0446
teams*zoom                  3     245.1250000      81.7083333       6.45    0.0045


Source                     DF     Type III SS     Mean Square    F Value    Pr > F

teams                       1     477.0416667     477.0416667      37.66    <.0001
zoom                        3     128.1250000      42.7083333       3.37    0.0446
teams*zoom                  3     245.1250000      81.7083333       6.45    0.0045
```

Least Squares Means
Adjustment for Multiple Comparisons: Bonferroni

| teams | zoom | response LSMEAN | LSMEAN Number |
|---|---|---|---|
| c | 0 | 93.0000000 | 1 |
| c | 100 | 92.0000000 | 2 |
| c | 150 | 91.0000000 | 3 |
| c | 50 | 94.0000000 | 4 |
| d | 0 | 77.0000000 | 5 |
| d | 100 | 84.0000000 | 6 |
| d | 150 | 92.0000000 | 7 |
| d | 50 | 81.3333333 | 8 |

Least Squares Means for effect teams*zoom
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: response

| i/j | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | | 1.0000 | 1.0000 | 1.0000 | **0.0013** | 0.1938 | 1.0000 | 0.0280 |
| 2 | 1.0000 | | 1.0000 | 1.0000 | 0.0026 | 0.3961 | 1.0000 | 0.0579 |
| 3 | 1.0000 | 1.0000 | | 1.0000 | 0.0053 | 0.7956 | 1.0000 | 0.1197 |
| 4 | 1.0000 | 1.0000 | 1.0000 | | 0.0007 | 0.0939 | 1.0000 | **0.0136** |
| 5 | 0.0013 | 0.0026 | 0.0053 | 0.0007 | | 0.7956 | **0.0026** | 1.0000 |
| 6 | 0.1938 | 0.3961 | 0.7956 | 0.0939 | 0.7956 | | 0.3961 | 1.0000 |
| 7 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.0026 | 0.3961 | | 0.0579 |
| 8 | 0.0280 | 0.0579 | 0.1197 | 0.0136 | 1.0000 | 1.0000 | 0.0579 | |

### Output Explanation (Part E. Bonferroni t Paired Comparisons)

The following pairs are significant because they have p-values less than the significance level 0.05: (5,1) (5,2) (5,3) (5,4) (7,5) (8,1) and (8,4). However, the pairs (5,2) (5,3) (5,4) and (8,1) are confounded pairs and do not contribute to the interaction. The significant differences in threat evaluation performance between unconfounded pairs in the Bonferroni analysis are the same as those found in the Newman-Keuls analysis. In the Williges (2006) reference, a fourth pair (7,8) is found to be significant and is not found to be significant in the SAS analysis. This p-value for the difference between 50% and 150% zoom displays used by distributed teams is just slightly larger (0.0579) than the stated p-value of 0.05. This occurs because in the SAS analysis the p-values are calculated for all pairs, not just unconfounded pairs.

## Example 20: One-Factor, Within-Subjects ANOVA

**(Click in this red rectangle to see the Reference Notes on Example 20.)**

### Problem

Location in Williges (2006) Table of Contents
Section 3, Topic 12. Within-Subjects ANOVA Designs, Part 12.1.1. Single Factor Design

Page(s) in Williges (2006) Reference Material: 405 – 408

Problem Description
Four enhancements using automated information to help soldiers work with battlefield information were evaluated. Four soldiers used each of four presentation enhancements (context dependent displays, intelligent tutors, multiple viewpoints, and groupware) to evaluate reconnaissance information for 35 different threats. Were the display enhancements significantly different ($p < 0.001$) in terms of the number of threats detected?

Context/Purpose
Determine if there is a significant difference among context dependent displays, intelligent tutors, multiple viewpoints, and groupware presentation enhancements in terms of the mean number of the 35 threats detected from the reconnaissance information.

Statistical Decision Criteria
Conduct a one-way, within-subjects ANOVA at the 0.001 level of significance. This is a within-subjects design because each of the four soldiers is exposed to each of the four presentation enhancements.

### SAS Input**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

***Note: The enhancement variable coding corresponds to the original data as: 1=context dependent, 2=intelligent tutors, 3=multiple view points, and 4=groupware.*

```
options nodate nocenter pageno=1;
title 'Example 20: One-Factor Within-Subjects';
data information;
input subject $ enhancement $ response;
lines;
1 1 14
2 1 9
3 1 19
4 1 19
1 2 18
2 2 15
3 2 21
4 2 18
1 3 18
2 3 17
3 3 26
```

```
4 3 21
1 4 20
2 4 19
3 4 30
4 4 27
;
proc glm;
class subject enhancement;
model response= subject enhancement subject*enhancement;
means subject enhancement/alpha=.001;
test h=enhancement e=subject*enhancement;
run;
quit;
```

## SAS Output

```
Example 20: One-Factor Within-Subjects

The GLM Procedure

      Class Level Information

Class            Levels    Values

subject               4    1 2 3 4

enhancement           4    1 2 3 4


Number of Observations Read          16
Number of Observations Used          16

Dependent Variable: response

                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F

Model                       15     387.9375000      25.8625000          .          .

Error                        0       0.0000000           .

Corrected Total             15     387.9375000


R-Square     Coeff Var      Root MSE    response Mean

1.000000           .              .         19.43750

Source                      DF       Type I SS     Mean Square    F Value    Pr > F

subject                      3     190.1875000      63.3958333          .          .
enhancement                  3     166.1875000      55.3958333          .          .
subject*enhancement          9      31.5625000       3.5069444          .          .
```

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **subject** | **3** | **190.1875000** | 63.3958333 | . | . |
| **enhancement** | **3** | **166.1875000** | **55.3958333** | . | . |
| **subject*enhancement** | **9** | **31.5625000** | **3.5069444** | . | . |

| Level of subject | N | response Mean | response Std Dev |
|---|---|---|---|
| 1 | 4 | 17.5000000 | 2.51661148 |
| 2 | 4 | 15.0000000 | 4.32049380 |
| 3 | 4 | 24.0000000 | 4.96655481 |
| 4 | 4 | 21.2500000 | 4.03112887 |

| Level of enhancement | N | response Mean | response Std Dev |
|---|---|---|---|
| 1 | 4 | 15.2500000 | 4.78713554 |
| 2 | 4 | 18.0000000 | 2.44948974 |
| 3 | 4 | 20.5000000 | 4.04145188 |
| 4 | 4 | 24.0000000 | 5.35412613 |

Dependent Variable: response

Tests of Hypotheses Using the Type III MS for subject*enhancement as an Error Term

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **enhancement** | **3** | **166.1875000** | **55.3958333** | **15.80** | **0.0006** |

### Output Explanation

Presentation enhancement had a significant effect on threat evaluation since the p-value (0.0006) is less than 0.001. The result is a significant effect on the information evaluations due to the presentation enhancements. Post hoc analyses are needed to determine which of the four types of presentation enhancements significantly affected the soldiers' evaluations.

## Example 21: Two-Factor, Within-Subjects ANOVA

**(Click in this red rectangle to see the Reference Notes on Example 21.)**

### Problem

Location in Williges (2006) Table of Contents
Section 3, Topic 12. Within-Subjects ANOVA Design, Part 12.1.2. Two-Factor Design

Page(s) in Williges (2006) Reference Material: 413 – 415

Problem Description
Three alternative visual displays (3 dimensional graphs, color coded diagrams, and flowcharts) were developed to augment intelligence information gathered over a 12-hour period. Six intelligence officers evaluated the information using each visual display either as redundant to or as a substitute for the standard intelligence information. Are the information presentations significantly different ($p < 0.05$)?

Context/Purpose
Determine if the main effect of three visual displays and two uses of displayed intelligence information are significantly different. In addition, determine if display type and display use interact significantly.

Statistical Decision Criteria
Since each officer is exposed to all three types of display and both uses of displayed information. The experimenter needs to perform a 3x2 within-subjects ANOVA at $\alpha = 0.05$.

### SAS Input***

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

*** Note: The symbol (|) used in the model statement indicates to SAS that the two variables surrounding it should be analyzed as main effects and interaction effects.*
*For this example response = use|display|subject is a compact way to write response = use display subject use*display use*subject display*subject use*display*subject.*
****Note: The coding of display corresponds to the original problem as: 1=three-dimensional graphs, 2=color-coded diagrams, and 3=flowcharts.*

```
options nodate nocenter pageno=1;
title 'Example 21: Two-Factors, Within-Subjects';
data six;
input subject $ use $ display $ response;
lines;
1 r 1 46
2 r 1 50
3 r 1 49
4 r 1 47
5 r 1 51
6 r 1 45
1 s 1 47
2 s 1 46
```

```
3 s 1 50
4 s 1 44
5 s 1 50
6 s 1 44
1 r 2 49
2 r 2 52
3 r 2 54
4 r 2 48
5 r 2 54
6 r 2 48
1 s 2 39
2 s 2 44
3 s 2 38
4 s 2 45
5 s 2 43
6 s 2 41
1 r 3 50
2 r 3 47
3 r 3 49
4 r 3 52
5 r 3 53
6 r 3 47
1 s 3 35
2 s 3 42
3 s 3 39
4 s 3 40
5 s 3 42
6 s 3 41
;
proc glm;
class use display subject;
model response = use|display|subject;
means use display use*display/alpha=0.05;
test h=use e=use*subject;
test h=display e=display*subject;
test h=use*display e=use*display*subject;
run;
quit;
```

### SAS Output

```
Example 21: Two-Factors, Within-Subjects
The GLM Procedure

      Class Level Information

Class         Levels    Values

use                2    r s

display            3    1 2 3

subject            6    1 2 3 4 5 6


Number of Observations Read          36
Number of Observations Used          36
```

Dependent Variable: response

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 35 | 800.3055556 | 22.8658730 | . | . |
| Error | 0 | 0.0000000 | . | | |
| **Corrected Total** | **35** | **800.3055556** | | | |

| R-Square | Coeff Var | Root MSE | response Mean |
|---|---|---|---|
| 1.000000 | . | . | 46.13889 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| use | 1 | 406.6944444 | 406.6944444 | . | . |
| display | 2 | 42.8888889 | 21.4444444 | . | . |
| use*display | 2 | 139.5555556 | 69.7777778 | . | . |
| subject | 5 | 86.4722222 | 17.2944444 | . | . |
| use*subject | 5 | 16.4722222 | 3.2944444 | . | . |
| display*subject | 10 | 31.4444444 | 3.1444444 | . | . |
| use*display*subject | 10 | 76.7777778 | 7.6777778 | . | . |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **use** | **1** | **406.6944444** | **406.6944444** | **.** | **.** |
| **display** | **2** | **42.8888889** | **21.4444444** | **.** | **.** |
| **use*display** | **2** | **139.5555556** | **69.7777778** | **.** | **.** |
| **subject** | **5** | **86.4722222** | **17.2944444** | **.** | **.** |
| **use*subject** | **5** | **16.4722222** | **3.2944444** | **.** | **.** |
| **display*subject** | **10** | **31.4444444** | **3.1444444** | **.** | **.** |
| **use*display*subject** | **10** | **76.7777778** | **7.6777778** | **.** | **.** |

| Level of use | N | response Mean | response Std Dev |
|---|---|---|---|
| r | 18 | 49.5000000 | 2.70620203 |
| s | 18 | 42.7777778 | 3.97870147 |

| Level of display | N | response Mean | response Std Dev |
|---|---|---|---|
| 1 | 12 | 47.4166667 | 2.50302847 |
| 2 | 12 | 46.2500000 | 5.49586622 |
| 3 | 12 | 44.7500000 | 5.69090183 |

| Level of use | Level of display | N | response Mean | response Std Dev |
|---|---|---|---|---|
| r | 1 | 6 | 48.0000000 | 2.36643191 |
| r | 2 | 6 | 50.8333333 | 2.85773803 |
| r | 3 | 6 | 49.6666667 | 2.50333111 |
| s | 1 | 6 | 46.8333333 | 2.71416040 |
| s | 2 | 6 | 41.6666667 | 2.80475786 |
| s | 3 | 6 | 39.8333333 | 2.63944439 |

Dependent Variable: response

Tests of Hypotheses Using the Type III MS for use*subject as an Error Term

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **use** | **1** | **406.6944444** | **406.6944444** | **123.45** | **0.0001** |

Tests of Hypotheses Using the Type III MS for display*subject as an Error Term

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **display** | **2** | **42.88888889** | **21.44444444** | **6.82** | **0.0135** |

Tests of Hypotheses Using the Type III MS for use*display*subject as an Error Term

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **use*display** | **2** | **139.5555556** | **69.7777778** | **9.09** | **0.0056** |

### Output Explanation

The main effects of information use, display type, and the use by display interaction are each significant because all three p-values are less than 0.05 (i.e., p = 0.0001, 0.0135, and 0.0056, respectively). Additional post hoc analyses are needed to resolve the many effects of display type and the use by display interaction because more than two treatment means are included in these effects.

## Example 22: Geisser-Greenhouse and Huyhn-Feldt Corrections

**(Click in this red rectangle to see the Reference Notes on Example 22.)**

### Problem

Location in Williges (2006) Table of Contents
Section 3, Topic 12. Within-Subjects ANOVA Design, Part 12.2. Homogeneity of Covariance

Page(s) in Williges (2006) Reference Material: 421 – 422

Problem Description
Part A: Four enhancements using automated information to help soldiers work with battlefield information were evaluated. Four soldiers used each of four presentation enhancements (context dependent displays, intelligent tutors, multiple viewpoints, and groupware) to evaluate reconnaissance information for 35 different threats. Were the display enhancements significantly different ($p < 0.001$) in terms of the number of threats detected?

Part B: Three alternative visual displays (3 dimensional graphs, color coded diagrams, and flowcharts) were developed to augment intelligence information gathered over a 12-hour period. Six intelligence officers evaluated the information using each visual display either as redundant to or as a substitute for the standard intelligence information. Are the information presentations significantly different ($p < 0.05$)?

Context/Purpose
Determine the extent of the Giesser-Greenhouse (G-G) and Huynh-Feldt (H-F) corrections for homogeneity of covariance for the problems described in Parts A and B.

Statistical Decision Criteria
Recalculate the ANOVAs on both the one and two factor within-subjects designs described in problems A and B to determine the G-G and H-F corrected p-values.

### SAS Input (Part A. One-Way, Within-Subjects ANOVA Correction)**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

***Note: This is the same problem and data used in Example 20. The form used in the input statement corresponds to the threats detected at each level of treatment enhancement. There is no class statement because there is no independent variable in the data input. The repeated command, allows SAS to calculate the repeated measures ANOVA. The dependent variables are included in the model statement, but since there is not a class statement, the area to the right of model statement is left empty. The nouni command indicates to SAS not to conduct univariate analyses on the dependent variables. See Cody and Smith (1997) for a more complete explanation of this format.*

```
options nodate nocenter pageno=1;
title 'Example 22A: Geisser-Greenhouse and Huynh-Feldt One-Way Corrections';
data within;
input threat1-threat4;
lines;
14 18 18 20
```

```
9 15 17 19
19 21 26 30
19 18 21 27
;
proc glm;
model threat1-threat4 = /nouni;
repeated enhancement 4 (1 2 3 4)/ printe;
run;
proc corr data=within;
run;
quit;
```

### SAS Output (Part A. One-Way, Within-Subjects ANOVA Correction)

```
Example 22A: Geisser-Greenhouse and Huynh-Feldt One-Way Corrections
1
The GLM Procedure

Number of Observations Read           4
Number of Observations Used           4


Repeated Measures Analysis of Variance

            Repeated Measures Level Information

  Dependent Variable     threat1  threat2  threat3  threat4

Level of enhancement           1        2        3        4


Partial Correlation Coefficients from the Error SSCP Matrix / Prob > |r|

DF = 3         threat1        threat2        threat3        threat4

threat1       1.000000       0.852803       0.818388       0.910359
                             0.1472         0.1816         0.0896

threat2       0.852803       1.000000       0.909137       0.838742
                0.1472                       0.0909         0.1613

threat3       0.818388       0.909137       1.000000       0.955090
                0.1816         0.0909                       0.0449

threat4       0.910359       0.838742       0.955090       1.000000
                0.0896         0.1613         0.0449

                                 E = Error SSCP Matrix

enhancement_N represents the contrast between the nth level of enhancement and the last

                   enhancement_1      enhancement_2      enhancement_3

enhancement_1              14.75              13.00               1.50
enhancement_2              13.00              38.00              18.00
enhancement_3               1.50              18.00              11.00
```

Partial Correlation Coefficients from the Error SSCP Matrix of the Variables Defined by the Specified Transformation / Prob > |r|

| DF = 3 | enhancement_1 | enhancement_2 | enhancement_3 |
|---|---|---|---|
| enhancement_1 | 1.000000 | 0.549105<br>0.4509 | 0.117760<br>0.8822 |
| enhancement_2 | 0.549105<br>0.4509 | 1.000000 | 0.880409<br>0.1196 |
| enhancement_3 | 0.117760<br>0.8822 | 0.880409<br>0.1196 | 1.000000 |

Repeated Measures Analysis of Variance

Sphericity Tests

| Variables | DF | Mauchly's Criterion | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| Transformed Variates | 5 | 0.0150067 | 7.2320544 | 0.2039 |
| Orthogonal Components | 5 | 0.0309138 | 5.9873974 | 0.3074 |

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no enhancement Effect
H = Type III SSCP Matrix for enhancement
E = Error SSCP Matrix

S=1 M=0.5 N=-0.5

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---|---|---|---|---|---|
| Wilks' Lambda | 0.00588428 | 56.31 | 3 | 1 | 0.0976 |
| Pillai's Trace | 0.99411572 | 56.31 | 3 | 1 | 0.0976 |
| Hotelling-Lawley Trace | 168.94444444 | 56.31 | 3 | 1 | 0.0976 |
| Roy's Greatest Root | 168.94444444 | 56.31 | 3 | 1 | 0.0976 |

```
Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

                                                                                  Adj Pr > F
Source                       DF     Type III SS     Mean Square    F Value    Pr > F     G - G     H - F

enhancement                   3     166.1875000      55.3958333      15.80    0.0006    0.0053    0.0006
Error(enhancement)            9      31.5625000       3.5069444


Greenhouse-Geisser Epsilon    0.6199
Huynh-Feldt Epsilon           1.5897


Example 22A: Geisser-Greenhouse and Huynh-Feldt, Within-Subjects
The CORR Procedure

   4  Variables:    threat1  threat2  threat3  threat4


                                  Simple Statistics

Variable           N          Mean       Std Dev           Sum       Minimum       Maximum

threat1            4      15.25000       4.78714      61.00000       9.00000      19.00000
threat2            4      18.00000       2.44949      72.00000      15.00000      21.00000
threat3            4      20.50000       4.04145      82.00000      17.00000      26.00000
threat4            4      24.00000       5.35413      96.00000      19.00000      30.00000


           Pearson Correlation Coefficients, N = 4
                 Prob > |r| under H0: Rho=0

              threat1       threat2       threat3       threat4

threat1       1.00000       0.85280       0.81839       0.91036
                             0.1472        0.1816        0.0896

threat2       0.85280       1.00000       0.90914       0.83874
               0.1472                      0.0909        0.1613

threat3       0.81839       0.90914       1.00000       0.95509
               0.1816        0.0909                      0.0449

threat4       0.91036       0.83874       0.95509       1.00000
               0.0896        0.1613        0.0449
```

### Output Explanation (Part A. One-Way, Within-Subjects ANOVA Correction)

The corrected G-G p-value (0.0053) and the corrected H-F p-value (0.0006) are quite similar to the uncorrected p = 0.0006. Each shows a significant enhancement main effect ($p < 0.001$). Likewise, the inter-correlations among the four levels of enhancement are quite similar suggesting sphericity or homogeneity of covariance. Note that the maximum G-G correction is slightly greater than the H-F correction which equals the uncorrected p-level.

### SAS Input (Part B. Two-Way, Within-Subjects ANOVA Correction)**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

***Note: This is the same problem and data used in Example 21. The form used in the input statement corresponds to the evaluations at each level of treatments alternative and use. There is no class statement because there is no independent variable in the data input. The repeated command, allows SAS to calculate the repeated measures ANOVA. The dependent variables are included in the model statement, but since there is not a class statement, the area to the right of model statement is left empty. The nouni command indicates to SAS not to conduct univariate analyses on the dependent variables. The nom command tells SAS to only display the univariate analyses. See Cody and Smith (1997) for a more complete explanation of this format.*

```
options nodate nocenter pageno=1;
title 'Example 22B: Geisser-Greenhouse and Huynh-Feldt Two-Way Corrections';
data six;
input evaluation1-evaluation6;
lines;
46 47 49 39 50 35
50 46 52 44 47 42
49 50 54 38 49 39
47 44 48 45 52 40
51 50 54 43 53 42
45 44 48 41 47 41
;
proc glm;
model evaluation1-evaluation6 = / nouni;
repeated alternative 3, use 2 /nom printe;
run;
proc corr data=six;
run;
quit;
```

### SAS Output (Part B. Two-Way, Within-Subjects ANOVA Correction)

```
Example 22B: Geisser-Greenhouse and Huynh-Feldt Two-Way Corrections

The GLM Procedure

Number of Observations Read          6
Number of Observations Used          6
```

Repeated Measures Analysis of Variance

Repeated Measures Level Information

| Dependent Variable | evaluation1 | evaluation2 | evaluation3 | evaluation4 | evaluation5 | evaluation6 |
|---|---|---|---|---|---|---|
| Level of alternative | 1 | 1 | 2 | 2 | 3 | 3 |
| Level of use | 1 | 2 | 1 | 2 | 1 | 2 |

Partial Correlation Coefficients from the Error SSCP Matrix / Prob > |r|

| DF = 5 | evaluation1 | evaluation2 | evaluation3 | evaluation4 | evaluation5 | evaluation6 |
|---|---|---|---|---|---|---|
| evaluation1 | 1.000000 | 0.685051<br>0.1332 | 0.887227<br>0.0184 | 0.271196<br>0.6032 | 0.303851<br>0.5583 | 0.512323<br>0.2988 |
| evaluation2 | 0.685051<br>0.1332 | 1.000000 | 0.898188<br>0.0150 | -0.455388<br>0.3641 | 0.343418<br>0.5051 | -0.060489<br>0.9094 |
| evaluation3 | 0.887227<br>0.0184 | 0.898188<br>0.0150 | 1.000000 | -0.182984<br>0.7286 | 0.158423<br>0.7644 | 0.313763<br>0.5448 |
| evaluation4 | 0.271196<br>0.6032 | -0.455388<br>0.3641 | -0.182984<br>0.7286 | 1.000000 | 0.265860<br>0.6106 | 0.639380<br>0.1716 |
| evaluation5 | 0.303851<br>0.5583 | 0.343418<br>0.5051 | 0.158423<br>0.7644 | 0.265860<br>0.6106 | 1.000000 | -0.070628<br>0.8942 |
| evaluation6 | 0.512323<br>0.2988 | -0.060489<br>0.9094 | 0.313763<br>0.5448 | 0.639380<br>0.1716 | -0.070628<br>0.8942 | 1.000000 |

E = Error SSCP Matrix

alternative_N represents the contrast between the nth level of alternative and the last

| | alternative_1 | alternative_2 |
|---|---|---|
| alternative_1 | 101.33 | 33.00 |
| alternative_2 | 33.00 | 26.00 |

Repeated Measures Analysis of Variance

Partial Correlation Coefficients from the Error SSCP Matrix of the Variables Defined by the Specified Transformation / Prob > |r|

| DF = 5 | alternative_1 | alternative_2 |
|---|---|---|
| alternative_1 | 1.000000 | 0.642911<br>0.1685 |
| alternative_2 | 0.642911<br>0.1685 | 1.000000 |

Sphericity Tests

| Variables | DF | Mauchly's Criterion | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| Transformed Variates | 2 | 0.3813218 | 3.8564467 | 0.1454 |
| Orthogonal Components | 2 | 0.5210828 | 2.6073853 | 0.2715 |

E = Error SSCP Matrix

use_N represents the contrast between the nth level of use and the last

| | use_1 |
|---|---|
| use_1 | 98.833 |

E = Error SSCP Matrix

alternative_N represents the contrast between the nth level of alternative and the last
use_N represents the contrast between the nth level of use and the last

| | alternative_1*use_1 | alternative_2*use_1 |
|---|---|---|
| alternative_1*use_1 | 133.333 | 52.333 |
| alternative_2*use_1 | 52.333 | 149.333 |

Repeated Measures Analysis of Variance

Partial Correlation Coefficients from the Error SSCP Matrix of the Variables Defined by the Specified Transformation / Prob > |r|

| DF = 5 | alternative_1*use_1 | alternative_2*use_1 |
|---|---|---|
| alternative_1*use_1 | 1.000000 | 0.370878<br>0.4692 |
| alternative_2*use_1 | 0.370878<br>0.4692 | 1.000000 |

Sphericity Tests

| Variables | DF | Mauchly's Criterion | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| Transformed Variates | 2 | 0.8596865 | 0.60475 | 0.7391 |
| Orthogonal Components | 2 | 0.9710397 | 0.1175518 | 0.9429 |

Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

| Source | DF | Type III SS | Mean Square | F Value | Pr > F | Adj Pr > F G - G | Adj Pr > F H - F |
|---|---|---|---|---|---|---|---|
| alternative | 2 | 42.88888889 | 21.44444444 | **6.82** | **0.0135** | **0.0303** | **0.0202** |
| Error(alternative) | 10 | 31.44444444 | 3.14444444 | | | | |

Greenhouse-Geisser Epsilon 0.6762
Huynh-Feldt Epsilon 0.8381

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| use | 1 | 406.6944444 | 406.6944444 | 123.45 | 0.0001 |
| Error(use) | 5 | 16.4722222 | 3.2944444 | | |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F | Adj Pr > F G - G | Adj Pr > F H - F |
|---|---|---|---|---|---|---|---|
| alternative*use | 2 | 139.5555556 | 69.7777778 | **9.09** | **0.0056** | **0.0062** | **0.0056** |
| Error(alternative*use) | 10 | 76.7777778 | 7.6777778 | | | | |

Greenhouse-Geisser Epsilon 0.9719
Huynh-Feldt Epsilon 1.5807

Example 22B: Geisser-Greenhouse and Huynh-Feldt, Within-Subjects
The CORR Procedure

6 Variables: evaluation1 evaluation2 evaluation3 evaluation4 evaluation5 evaluation6

Simple Statistics

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| evaluation1 | 6 | 48.00000 | 2.36643 | 288.00000 | 45.00000 | 51.00000 |
| evaluation2 | 6 | 46.83333 | 2.71416 | 281.00000 | 44.00000 | 50.00000 |
| evaluation3 | 6 | 50.83333 | 2.85774 | 305.00000 | 48.00000 | 54.00000 |
| evaluation4 | 6 | 41.66667 | 2.80476 | 250.00000 | 38.00000 | 45.00000 |
| evaluation5 | 6 | 49.66667 | 2.50333 | 298.00000 | 47.00000 | 53.00000 |
| evaluation6 | 6 | 39.83333 | 2.63944 | 239.00000 | 35.00000 | 42.00000 |

Pearson Correlation Coefficients, N = 6
Prob > |r| under H0: Rho=0

| | evaluation1 | evaluation2 | evaluation3 | evaluation4 | evaluation5 | evaluation6 |
|---|---|---|---|---|---|---|
| evaluation1 | 1.00000 | 0.68505 | 0.88723 | 0.27120 | 0.30385 | 0.51232 |
| | | 0.1332 | 0.0184 | 0.6032 | 0.5583 | 0.2988 |
| evaluation2 | 0.68505 | 1.00000 | 0.89819 | -0.45539 | 0.34342 | -0.06049 |
| | 0.1332 | | 0.0150 | 0.3641 | 0.5051 | 0.9094 |
| evaluation3 | 0.88723 | 0.89819 | 1.00000 | -0.18298 | 0.15842 | 0.31376 |
| | 0.0184 | 0.0150 | | 0.7286 | 0.7644 | 0.5448 |
| evaluation4 | 0.27120 | -0.45539 | -0.18298 | 1.00000 | 0.26586 | 0.63938 |
| | 0.6032 | 0.3641 | 0.7286 | | 0.6106 | 0.1716 |
| evaluation5 | 0.30385 | 0.34342 | 0.15842 | 0.26586 | 1.00000 | -0.07063 |
| | 0.5583 | 0.5051 | 0.7644 | 0.6106 | | 0.8942 |
| evaluation6 | 0.51232 | -0.06049 | 0.31376 | 0.63938 | -0.07063 | 1.00000 |
| | 0.2988 | 0.9094 | 0.5448 | 0.1716 | 0.8942 | |

**Output Explanation (Part B. Two-Way, Within-Subjects ANOVA Correction)**

The SAS analysis does not calculate the corrected p-values for the main effect of Use because it has only two levels. However, the main effect of Use is significant as shown in Williges (2006). All the G-G and H-F corrected p-values and the uncorrected p-levels are significant at the 0.05 level for the main effect of Alternative and the two-way interaction. Corrections to p-levels are greatest for the interaction where the inter-correlations among the six treatment levels range from .0150 to 0.9094 suggesting some degree of heterogeneity of covariance. Again note that the maximum G-G correction is slightly greater than the H-F p-level correction as expected.

## Example 23: Testing Order Effects in Balanced Latin Squares

**(Click in this red rectangle to see the Reference Notes on Example 23.)**

### Problem

Location in Williges (2006) Table of Contents
Section 3, Topic 12. Within-Subjects ANOVA Design, Part 12.3.3. Testing Order Effects

Page(s) in Williges (2006) Reference Material: 432 – 436

Problem Description
Four enhancements using automated information to help soldiers work with battlefield information were evaluated. Four soldiers used each of four presentation enhancements (context dependent displays, intelligent tutors, multiple viewpoints, and groupware) to evaluate reconnaissance information for 35 different threats. Was the effect of presentation order of the four treatments significantly different ($p < 0.001$)?

Context/Purpose
A 4x4 Balanced Latin Square was used to counterbalance the order and partially balance the sequence of presentation of the four enhancement alternatives to each of the four soldiers. Can the presentation order effect be significant even though it was balanced across treatments and independent of the treatment effect?

Statistical Decision Criteria
Conduct an ANOVA on the Balanced Latin Square used for treatment presentation order to determine if there is a significant order effect at the 0.001 level of significance.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

***Note: The data presented here is the same data that is used in Examples 20 and 22. The coding for the enhancement variable is the same as in Example 20.*

```
options nodate nocenter pageno=1;
title 'Example 23: Testing Order Effects in Balanced Latin Squares';
data information;
input order $ subject $ enhancement $ response;
lines;
1 1 1 14
1 2 2 15
1 3 3 26
1 4 4 27
2 1 2 18
2 2 3 17
2 3 4 30
2 4 1 19
3 1 4 20
3 2 1 9
3 3 2 21
3 4 3 21
```

```
4 1 3 18
4 2 4 19
4 3 1 19
4 4 2 18
;
proc glm;
class subject order enhancement;
model response = subject order enhancement;
means subject order enhancement/alpha=0.001;
run;
quit;
```

### SAS Output**

***Note: The output results are slightly different than those in the Williges (2006) reference due to rounding.*

```
Example 23: Testing Order Effects in Balanced Latin Squares

The GLM Procedure

      Class Level Information

Class            Levels    Values

subject               4    1 2 3 4

order                 4    1 2 3 4

enhancement           4    1 2 3 4


Number of Observations Read          16
Number of Observations Used          16
```

Dependent Variable: response

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 9 | 385.5625000 | 42.8402778 | 108.23 | <.0001 |
| **Error** | **6** | **2.3750000** | **0.3958333** | | |
| **Corrected Total** | **15** | **387.9375000** | | | |

| R-Square | Coeff Var | Root MSE | response Mean |
|---|---|---|---|
| 0.993878 | 3.236799 | 0.629153 | 19.43750 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| subject | 3 | 190.1875000 | 63.3958333 | 160.16 | <.0001 |
| order | 3 | 29.1875000 | 9.7291667 | 24.58 | 0.0009 |
| enhancement | 3 | 166.1875000 | 55.3958333 | 139.95 | <.0001 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **subject** | **3** | **190.1875000** | **63.3958333** | **160.16** | **<.0001** |
| **order** | **3** | **29.1875000** | **9.7291667** | **24.58** | **0.0009** |
| **enhancement** | **3** | **166.1875000** | **55.3958333** | **139.95** | **<.0001** |

| Level of subject | N | response Mean | response Std Dev |
|---|---|---|---|
| 1 | 4 | 17.5000000 | 2.51661148 |
| 2 | 4 | 15.0000000 | 4.32049380 |
| 3 | 4 | 24.0000000 | 4.96655481 |
| 4 | 4 | 21.2500000 | 4.03112887 |

| Level of order | N | response Mean | response Std Dev |
|---|---|---|---|
| 1 | 4 | **20.5000000** | 6.95221787 |
| 2 | 4 | **21.0000000** | 6.05530071 |
| 3 | 4 | **17.7500000** | 5.85234996 |
| 4 | 4 | **18.5000000** | 0.57735027 |

| Level of enhancement | N | response Mean | response Std Dev |
|---|---|---|---|
| 1 | 4 | 15.2500000 | 4.78713554 |
| 2 | 4 | 18.0000000 | 2.44948974 |
| 3 | 4 | 20.5000000 | 4.04145188 |
| 4 | 4 | 24.0000000 | 5.35412613 |

**Output Explanation**

This ANOVA on the Balanced Latin Square resulted in a significant effect due to the presentation order of the enhancements since the p-value (0.0009) is less than 0.001. This effect is independent of the significant treatment and subject effects. Consequently, the Balanced Latin Square procedure for partially counterbalancing order and sequence effects was successful in keeping the confounding effect of presentation order independent of the treatment effect of interest to the experiment.

## Example 24: Within-Subjects and Between-Subjects Design Comparison

**(Click in this red rectangle to see the Reference Notes on Example 24.)**

### Problem

Location in Williges (2006) Table of Contents
Section 3, Topic 12. Within-Subjects ANOVA Design, Part 12.5. Within-Subjects Design Advantages

Page(s) in Williges (2006) Reference Material: 439

Problem Description
Four enhancements using automated information to help soldiers work with battlefield information were evaluated. Four soldiers used each of four presentation enhancements (context dependent displays, intelligent tutors, multiple viewpoints, and groupware) to evaluate reconnaissance information for 35 different threats. Were the display enhancements significantly different ($p < 0.001$) in terms of the number of threats detected?

Context/Purpose
Compare the sensitivity of using a within-subjects design to its between-subjects design alternative.

Statistical Decision Criteria
Perform both a within-subjects and between-subjects ANOVA to test the significance differences ($p < 0.001$) among the four presentation enhancements.

### SAS Input (Part A. Within-Subjects ANOVA)**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

***Note: This is the same within-subjects analysis performed in Example 20.*

```
options nodate nocenter pageno=1;
title 'Example 24A: One-Factor Within-Subjects';
data information;
input subject $ enhancement $ response;
lines;
1 1 14
2 1 9
3 1 19
4 1 19
1 2 18
2 2 15
3 2 21
4 2 18
1 3 18
2 3 17
3 3 26
4 3 21
1 4 20
```

```
2 4 19
3 4 30
4 4 27
;
proc glm;
class subject enhancement;
model response= subject enhancement subject*enhancement;
means subject enhancement/alpha=.001;
test h=enhancement e=subject*enhancement;
run;
quit;
```

### SAS Output (Part A. Within-Subjects ANOVA)

```
Example 24A: One-Factor Within-Subjects

The GLM Procedure

      Class Level Information

Class            Levels    Values

subject               4    1 2 3 4

enhancement           4    1 2 3 4


Number of Observations Read          16
Number of Observations Used          16


Dependent Variable: response

                                   Sum of
Source                     DF         Squares     Mean Square    F Value    Pr > F

Model                      15     387.9375000      25.8625000        .       .

Error                       0       0.0000000        .

Corrected Total            15     387.9375000


R-Square     Coeff Var      Root MSE    response Mean

1.000000           .              .         19.43750


Source                     DF       Type I SS     Mean Square    F Value    Pr > F

subject                     3     190.1875000      63.3958333        .       .
enhancement                 3     166.1875000      55.3958333        .       .
subject*enhancement         9      31.5625000       3.5069444        .       .
```

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **subject** | **3** | **190.1875000** | **63.3958333** | . | . |
| **enhancement** | **3** | **166.1875000** | **55.3958333** | . | . |
| **subject*enhancement** | **9** | **31.5625000** | **3.5069444** | . | . |

| Level of subject | N | response Mean | response Std Dev |
|---|---|---|---|
| 1 | 4 | 17.5000000 | 2.51661148 |
| 2 | 4 | 15.0000000 | 4.32049380 |
| 3 | 4 | 24.0000000 | 4.96655481 |
| 4 | 4 | 21.2500000 | 4.03112887 |

| Level of enhancement | N | response Mean | response Std Dev |
|---|---|---|---|
| 1 | 4 | 15.2500000 | 4.78713554 |
| 2 | 4 | 18.0000000 | 2.44948974 |
| 3 | 4 | 20.5000000 | 4.04145188 |
| 4 | 4 | 24.0000000 | 5.35412613 |

Dependent Variable: response

Tests of Hypotheses Using the Type III MS for subject*enhancement as an Error Term

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **enhancement** | **3** | **166.1875000** | **55.3958333** | **15.80** | **0.0006** |

### Output Explanation (Part A. Within-Subjects ANOVA)

Presentation enhancement is significant at p = 0.0006.

### SAS Input (Part B. Between-Subjects ANOVA)**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

***Note: This is the same data as used in part A and Example 20, but it has been modified to be a between-subjects design.*

```
options nodate nocenter pageno=1;
title 'Example 24B: One-Factor Between-Subjects';
data information;
input subject $ enhancement $ response;
lines;
1 1 14
2 1 9
3 1 19
4 1 19
5 2 18
6 2 15
7 2 21
8 2 18
9 3 18
10 3 17
11 3 26
12 3 21
13 4 20
14 4 19
15 4 30
16 4 27
;
proc glm;
class subject enhancement;
model response= enhancement subject(enhancement);
means enhancement/alpha=.001;
test h=enhancement e=subject(enhancement);
run;
quit;
```

### SAS Output (Part B. Between-Subjects ANOVA)

```
Example 24B: One-Factor Between-Subjects

The GLM Procedure

                    Class Level Information

Class            Levels    Values

subject              16    1 10 11 12 13 14 15 16 2 3 4 5 6 7 8 9

enhancement           4    1 2 3 4

Number of Observations Read          16
Number of Observations Used          16
```

```
Dependent Variable: response

                                     Sum of
Source                       DF         Squares     Mean Square    F Value    Pr > F

Model                        15     387.9375000      25.8625000        .       .

Error                         0       0.0000000        .

Corrected Total              15     387.9375000


R-Square     Coeff Var      Root MSE    response Mean

1.000000           .              .         19.43750


Source                       DF       Type I SS     Mean Square    F Value    Pr > F

enhancement                   3     166.1875000      55.3958333        .       .
subject(enhancement)         12     221.7500000      18.4791667        .       .


Source                       DF     Type III SS     Mean Square    F Value    Pr > F

enhancement                   3     166.1875000      55.3958333        .       .
subject(enhancement)         12     221.7500000      18.4791667        .       .

Level of              -----------response----------
enhancement     N             Mean          Std Dev

1               4       15.2500000       4.78713554
2               4       18.0000000       2.44948974
3               4       20.5000000       4.04145188
4               4       24.0000000       5.35412613


Dependent Variable: response

Tests of Hypotheses Using the Type III MS for subject(enhancement) as an Error Term

Source                       DF     Type III SS     Mean Square    F Value    Pr > F

enhancement                   3     166.1875000      55.3958333       3.00    0.0729
```

### Output Explanation (Part B. Between-Subjects ANOVA)

The p-level (0.0729) of the main effect of presentation enhancement is not significant at the 0.001 level in the between-subjects ANOVA. By comparison, the alternative within-subjects design test of the presentation enhancement ANOVA main effect was significant at p = 0.0006 as shown in Part A. These analyses illustrated that the within-subjects design provides a more sensitive (powerful) F-test than its between-subjects counterpart.

## Example 25: Two-Way, Mixed-Factors ANOVA

**(Click in this red rectangle to see the Reference Notes on Example 25.)**

### Problem

Location in Williges (2006) Table of Contents
Section 3, Topic 13. Mixed-Factors ANOVA Designs, Part 13.1.2. Two-Factor Design Example

Page(s) in Williges (2006) Reference Material: 448 – 452

Problem Description
The decrement in target detection across 1-hour monitoring sessions was measured every 20 minutes for five soldiers who monitored displays where the ratio of targets to non-targets was either 9/1 or 1/9. Are there any significant effects ($p < 0.05$) in the percent of defined targets detected in this experiment?

Context/Purpose
Determine if there are significant differences in target detection due to time monitoring, the ratio of targets to non-targets, or the interaction of time monitoring and target ratios.

Statistical Decision Criteria
Conduct a 2x3 mixed-factors ANOVA to determine if there are significant effects of time or ratio at the 0.05 level of significance. This is a mixed-factors design because time is a between-subjects factor and ratio is a within-subjects factor.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 25: Two-Way, Mixed-Factors ANOVA';
data detection;
input subject $ ratio $ time $ targets;
lines;
1 1/9 20 95
1 1/9 40 90
1 1/9 60 82
2 1/9 20 89
2 1/9 40 82
2 1/9 60 83
3 1/9 20 92
3 1/9 40 80
3 1/9 60 79
4 1/9 20 86
4 1/9 40 89
4 1/9 60 77
5 1/9 20 90
5 1/9 40 92
5 1/9 60 75
6 9/1 20 90
6 9/1 40 88
```

```
6 9/1 60 92
7 9/1 20 87
7 9/1 40 95
7 9/1 60 95
8 9/1 20 96
8 9/1 40 93
8 9/1 60 95
9 9/1 20 94
9 9/1 40 90
9 9/1 60 90
10 9/1 20 91
10 9/1 40 87
10 9/1 60 86
;
proc glm;
class subject ratio time;
model targets = ratio time subject(ratio) ratio*time time*subject(ratio);
means ratio time ratio*time/alpha=0.05;
test h=ratio e=subject(ratio);
test h=time e=time*subject(ratio);
test h=ratio*time e=time*subject(ratio);
run;
quit;
```

### SAS Output

```
Example 25: Two-Way, Mixed-Factors ANOVA

The GLM Procedure

           Class Level Information

Class         Levels    Values

subject           10    1 10 2 3 4 5 6 7 8 9

ratio              2    1/9 9/1

time               3    20 40 60


Number of Observations Read          30
Number of Observations Used          30
```

Dependent Variable: targets

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 29 | 938.6666667 | 32.3678161 | . | . |
| Error | 0 | 0.0000000 | . | | |
| **Corrected Total** | **29** | **938.6666667** | | | |

| R-Square | Coeff Var | Root MSE | targets Mean |
|---|---|---|---|
| 1.000000 | . | . | 88.33333 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| ratio | 1 | 258.1333333 | 258.1333333 | . | . |
| time | 2 | 157.8666667 | 78.9333333 | . | . |
| subject(ratio) | 8 | 130.5333333 | 16.3166667 | . | . |
| ratio*time | 2 | 169.8666667 | 84.9333333 | . | . |
| subject*time(ratio) | 16 | 222.2666667 | 13.8916667 | . | . |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **ratio** | **1** | **258.1333333** | **258.1333333** | **.** | **.** |
| **time** | **2** | **157.8666667** | **78.9333333** | **.** | **.** |
| **subject(ratio)** | **8** | **130.5333333** | **16.3166667** | **.** | **.** |
| **ratio*time** | **2** | **169.8666667** | **84.9333333** | **.** | **.** |
| **subject*time(ratio)** | **16** | **222.2666667** | **13.8916667** | **.** | **.** |

```
Level of            -----------targets-----------
ratio         N             Mean          Std Dev

1/9          15       85.4000000       6.12722263
9/1          15       91.2666667       3.32665999


Level of            -----------targets-----------
time          N             Mean          Std Dev

20           10       91.0000000       3.29983165
40           10       88.6000000       4.67142614
60           10       85.4000000       7.35149267


Level of     Level of           -----------targets-----------
ratio        time         N             Mean          Std Dev

1/9          20           5       90.4000000       3.36154726
1/9          40           5       86.6000000       5.27257053
1/9          60           5       79.2000000       3.34664011
9/1          20           5       91.6000000       3.50713558
9/1          40           5       90.6000000       3.36154726
9/1          60           5       91.6000000       3.7815340


Dependent Variable: targets

   Tests of Hypotheses Using the Type III MS for subject(ratio) as an Error Term

Source                       DF     Type III SS     Mean Square    F Value    Pr > F

ratio                         1     258.1333333     258.1333333      15.82    0.0041


Tests of Hypotheses Using the Type III MS for subject*time(ratio) as an Error Term

Source                       DF     Type III SS     Mean Square    F Value    Pr > F

time                          2     157.8666667      78.9333333       5.68    0.0137
ratio*time                    2     169.8666667      84.9333333       6.11    0.0107
```

### Output Explanation

All three effects tested in this mixed-factors design are statistically significant, because the p-value for the test of the main effect of the ratio of targets to non-targets (0.0041), the main effect of time (0.0137), and the interaction between ratio and time (0.0107) are each less than the stated 0.05 significance level. The 9/1 ratio of targets to non-targets resulted in higher target detection than the 1/9 ratio. Further analyses are needed to interpret the significant effects of time monitoring, and the ratio by time interaction since more than two comparisons are involved in each interpretation.

# Section 4. Advanced ANOVA Designs

## Example 26: Complete Hierarchical Between-Subjects Design

**(Click in this red rectangle to see the Reference Notes on Example 26.)**

### Problem

Location in Williges (2006) Table of Contents
Section 4, Topic 16. Hierarchical Designs, Part 16.2.1. Complete Hierarchical Design

Page(s) in Williges (2006) Reference Material: 504-508

Problem Description
The military is testing a computer-based multimedia training procedure for commanders. The training procedure is presented to 80 commanders from eight battalions. Two battalions were chosen from each of two brigades within two divisions (infantry and cavalry). The hours to complete the multimedia training on the use of computer-generated surveillance displays were recorded for 10 commanders per battalion. Is training completion time significantly different based on the three command levels? ($p < 0.05$)

Context/Purpose
Determine if multimedia training completion time is significantly different for battalion commanders nested within brigades and divisions, brigades nested within divisions, and infantry and cavalry divisions.

Statistical Decision Criteria
Conduct a complete hierarchical ANOVA to test significant ($p < 0.05$) differences in training completion time across the level levels of command.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 26: Complete Hierarchical Between-Subjects Design';
data info;
input subject division brigade battalion hours;
lines;
1 1 1 1 17
2 1 1 1 28
3 1 1 1 16
4 1 1 1 13
5 1 1 1 31
6 1 1 1 27
7 1 1 1 23
8 1 1 1 16
9 1 1 1 34
10 1 1 1 12
11 1 1 2 29
12 1 1 2 35
13 1 1 2 33
14 1 1 2 29
15 1 1 2 37
16 1 1 2 25
```

```
17 1 1 2 32
18 1 1 2 13
19 1 1 2 26
20 1 1 2 29
21 1 2 3 34
22 1 2 3 23
23 1 2 3 34
24 1 2 3 33
25 1 2 3 18
26 1 2 3 26
27 1 2 3 12
28 1 2 3 27
29 1 2 3 24
30 1 2 3 19
31 1 2 4 39
32 1 2 4 21
33 1 2 4 10
34 1 2 4 18
35 1 2 4 23
36 1 2 4 17
37 1 2 4 34
38 1 2 4 39
39 1 2 4 33
40 1 2 4 35
41 2 3 5 23
42 2 3 5 17
43 2 3 5 36
44 2 3 5 21
45 2 3 5 12
46 2 3 5 28
47 2 3 5 32
48 2 3 5 24
49 2 3 5 17
50 2 3 5 36
51 2 3 6 13
52 2 3 6 24
53 2 3 6 11
54 2 3 6 19
55 2 3 6 20
56 2 3 6 33
57 2 3 6 22
58 2 3 6 14
59 2 3 6 19
60 2 3 6 36
61 2 4 7 15
62 2 4 7 25
63 2 4 7 30
64 2 4 7 32
65 2 4 7 40
66 2 4 7 28
67 2 4 7 33
68 2 4 7 16
69 2 4 7 39
70 2 4 7 32
71 2 4 8 15
72 2 4 8 27
73 2 4 8 25
```

```
74 2 4 8 18
75 2 4 8 20
76 2 4 8 28
77 2 4 8 11
78 2 4 8 22
79 2 4 8 13
80 2 4 8 25
;
proc glm;
class subject division brigade battalion;
model hours = division brigade(division) battalion(division brigade)
subject(division brigade battalion);
means division brigade(division) battalion(division brigade)/alpha=0.05;
test h=division e=subject(division brigade battalion);
test h=brigade(division) e=subject(division brigade battalion);
test h=battalion(division brigade) e=subject(division brigade battalion);
run;
quit;
```

### SAS Output

```
Example 26: Complete Hierarchical Between-Subjects Design                    1

The GLM Procedure

                                Class Level Information

Class        Levels  Values

subject          80  1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
                     29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53
                     54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78
                     79 80

division          2  1 2

brigade           4  1 2 3 4

battalion         8  1 2 3 4 5 6 7 8


Number of Observations Read          80
Number of Observations Used          80



Dependent Variable: hours

                                   Sum of
Source                     DF        Squares     Mean Square    F Value    Pr > F

Model                      79     5405.187500       68.420095        .       .

Error                       0        0.000000         .

Corrected Total            79     5405.187500
```

| R-Square | Coeff Var | Root MSE | hours Mean |
|---|---|---|---|
| 1.000000 | . | . | 24.68750 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| division | 1 | 66.612500 | 66.612500 | . | . |
| brigade(division) | 2 | 39.125000 | 19.562500 | . | . |
| batta(divisi*brigad) | 4 | 701.150000 | 175.287500 | . | . |
| subj(divi*brig*batt) | 72 | 4598.300000 | 63.865278 | . | . |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| division | 1 | 66.612500 | 66.612500 | . | . |
| brigade(division) | 2 | 39.125000 | 19.562500 | . | . |
| batta(divisi*brigad) | 4 | 701.150000 | 175.287500 | . | . |
| subj(divi*brig*batt) | 72 | 4598.300000 | 63.865278 | . | . |

| Level of division | hours N | hours Mean | hours Std Dev |
|---|---|---|---|
| 1 | 40 | 25.6000000 | 8.31063576 |
| 2 | 40 | 23.7750000 | 8.23528213 |

| Level of brigade | Level of division | hours N | hours Mean | hours Std Dev |
|---|---|---|---|---|
| 1 | 1 | 20 | 25.2500000 | 8.01889217 |
| 2 | 1 | 20 | 25.9500000 | 8.78680230 |
| 3 | 2 | 20 | 22.8500000 | 8.20317653 |
| 4 | 2 | 20 | 24.7000000 | 8.37351715 |

| Level of battalion | Level of division | Level of brigade | hours N | hours Mean | hours Std Dev |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 10 | 21.7000000 | 7.9169298 |
| 2 | 1 | 1 | 10 | 28.8000000 | 6.7131711 |
| 3 | 1 | 2 | 10 | 25.0000000 | 7.3786479 |
| 4 | 1 | 2 | 10 | 26.9000000 | 10.3220368 |
| 5 | 2 | 3 | 10 | 24.6000000 | 8.2758014 |
| 6 | 2 | 3 | 10 | 21.1000000 | 8.1710872 |
| 7 | 2 | 4 | 10 | 29.0000000 | 8.4195540 |
| 8 | 2 | 4 | 10 | 20.4000000 | 6.0037026 |

```
Dependent Variable: hours

Tests of Hypotheses Using the Type III MS for subj(divi*brig*batt) as an Error Term

Source                       DF     Type III SS     Mean Square    F Value    Pr > F

division                      1      66.6125000      66.6125000       1.04    0.3105
brigade(division)             2      39.1250000      19.5625000       0.31    0.7371
batta(divisi*brigad)          4     701.1500000     175.2875000       2.74    0.0348
```

### Output Explanation

Of the three hypothesis tests, only one is statistically significant. The test of the battalion commanders nested within divisions and brigades is significant since the p-value (0.035) is less than the stated significance level (0.05). Therefore, there is a significant difference in training completion time among battalion commanders nested within brigades and divisions. Post hoc tests are needed to isolate differences among battalion commanders. Possible interactions among battalion, brigade, and division command structure on multimedia training time completion cannot be assessed due to the complete nesting in this experimental design.

## Example 27: Partial Hierarchical Between-Subjects Design

**(Click in this red rectangle to see the Reference Notes on Example 27.)**

### Problem

Location in Williges (2006) Table of Contents
Section 4, Topic 16. Hierarchical Designs, Part 16.2.2. Partial Hierarchical Design

Page(s) in Williges (2006) Reference Material: 509-515

Problem Description
The military is testing two communication systems used by commanders of four brigades. Two brigades came from an infantry division and two from an armored division. Video conferencing and instant messaging are presented to 10 commanders in each brigade. Each commander used only one of the communication systems. The commanders' satisfaction ratings for the systems were recorded. Is there a significant satisfaction difference ($p < 0.05$) between the two communication systems and/or the nesting of commander levels?

Context/Purpose
Determine if there is a significant difference in the ratings of the two communication systems and command structure of battalion commanders.

Statistical Decision Criteria
A between-subjects, partial hierarchical ANOVA design is used to evaluate the satisfaction with communication systems where the two communication systems are crossed with battalion commanders nest within divisions, and with the infantry and armor divisions.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 27: Partial Hierarchical Between-Subjects Design';
data info;
input subject division brigade system hours;
lines;
1 1 1 1 17
2 1 1 1 28
3 1 1 1 16
4 1 1 1 13
5 1 1 1 21
6 1 1 1 27
7 1 1 1 23
8 1 1 1 16
9 1 1 1 23
10 1 1 1 12
11 1 2 1 29
12 1 2 1 35
13 1 2 1 33
14 1 2 1 29
15 1 2 1 37
16 1 2 1 25
```

```
17 1 2 1 32
18 1 2 1 13
19 1 2 1 26
20 1 2 1 29
21 2 3 1 34
22 2 3 1 23
23 2 3 1 39
24 2 3 1 33
25 2 3 1 19
26 2 3 1 26
27 2 3 1 12
28 2 3 1 27
29 2 3 1 24
30 2 3 1 19
31 2 4 1 39
32 2 4 1 21
33 2 4 1 10
34 2 4 1 18
35 2 4 1 23
36 2 4 1 17
37 2 4 1 34
38 2 4 1 39
39 2 4 1 29
40 2 4 1 35
41 1 1 2 23
42 1 1 2 17
43 1 1 2 36
44 1 1 2 21
45 1 1 2 12
46 1 1 2 28
47 1 1 2 32
48 1 1 2 24
49 1 1 2 17
50 1 1 2 20
51 1 2 2 13
52 1 2 2 24
53 1 2 2 11
54 1 2 2 19
55 1 2 2 20
56 1 2 2 33
57 1 2 2 22
58 1 2 2 14
59 1 2 2 19
60 1 2 2 36
61 2 3 2 15
62 2 3 2 25
63 2 3 2 30
64 2 3 2 32
65 2 3 2 40
66 2 3 2 28
67 2 3 2 33
68 2 3 2 16
69 2 3 2 39
70 2 3 2 32
71 2 4 2 35
72 2 4 2 27
73 2 4 2 35
```

```
74 2 4 2 18
75 2 4 2 20
76 2 4 2 28
77 2 4 2 11
78 2 4 2 22
79 2 4 2 13
80 2 4 2 25
;
proc glm;
class subject division brigade system;
model hours = division brigade(division) system system*division
system*brigade(division) subject(division brigade system);
lsmeans division brigade(division) system*division system*brigade(division);
test h=division e=subject(division brigade system);
test h=brigade(division)e=subject(division brigade system);
test h=system e=subject(division brigade system);
test h=system*division e=subject(division brigade system);
test h=system*brigade(division) e=subject(division brigade system);
run;
quit;
```

## SAS Output

```
Example 27: Partial Hierarchical Between-Subjects Design                                1

The GLM Procedure

                                Class Level Information

Class       Levels  Values

subject         80  1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
                    29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53
                    54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78
                    79 80

division         2  1 2

brigade          4  1 2 3 4

system           2  1 2


Number of Observations Read          80
Number of Observations Used          80
```

```
Dependent Variable: hours

                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F

Model                       79     5362.750000       67.882911        .          .

Error                        0        0.000000         .

Corrected Total             79     5362.750000


R-Square     Coeff Var      Root MSE    hours Mean

1.000000           .               .      24.62500


Source                      DF       Type I SS     Mean Square    F Value    Pr > F

division                     1      180.000000      180.000000        .          .
brigade(division)            2      188.450000       94.225000        .          .
system                       1       20.000000       20.000000        .          .
division*system              1       26.450000       26.450000        .          .
briga*system(divisi)         2      413.650000      206.825000        .          .
subj(divi*brig*syst)        72     4534.200000       62.975000        .          .


Source                      DF     Type III SS     Mean Square    F Value    Pr > F

division                     1      180.000000      180.000000        .          .
brigade(division)            2      188.450000       94.225000        .          .
system                       1       20.000000       20.000000        .          .
division*system              1       26.450000       26.450000        .          .
briga*system(divisi)         2      413.650000      206.825000        .          .
subj(divi*brig*syst)        72     4534.200000       62.975000        .          .

Least Squares Means

division    hours LSMEAN

1             23.1250000
2             26.1250000

brigade    division    hours LSMEAN

1          1             21.3000000
2          1             24.9500000
3          2             27.3000000
4          2             24.9500000

division    system    hours LSMEAN

1           1           24.2000000
1           2           22.0500000
2           1           26.0500000
2           2           26.2000000
```

| brigade | system | division | hours LSMEAN |
|---|---|---|---|
| 1 | 1 | 1 | 19.6000000 |
| 1 | 2 | 1 | 23.0000000 |
| 2 | 1 | 1 | 28.8000000 |
| 2 | 2 | 1 | 21.1000000 |
| 3 | 1 | 2 | 25.6000000 |
| 3 | 2 | 2 | 29.0000000 |
| 4 | 1 | 2 | 26.5000000 |
| 4 | 2 | 2 | 23.4000000 |

Dependent Variable: hours

Tests of Hypotheses Using the Type III MS for subj(divi*brig*syst) as an Error Term

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| division | 1 | 180.0000000 | 180.0000000 | 2.86 | 0.0952 |
| brigade(division) | 2 | 188.4500000 | 94.2250000 | 1.50 | 0.2309 |
| system | 1 | 20.0000000 | 20.0000000 | 0.32 | 0.5748 |
| division*system | 1 | 26.4500000 | 26.4500000 | 0.42 | 0.5190 |
| **briga*system(divisi)** | **2** | **413.6500000** | **206.8250000** | **3.28** | **0.0432** |

### Output Explanation

There is not a significant effect due to divisions, the nesting of brigades within divisions, communication systems, the communication system by brigades nested within divisions since the p-value (0.23) is greater than the stated significance level (0.05). The only significant effect is due to the interaction of communication systems and brigades nested with divisions since the p-value (0.043) is less than the stated significance level (0.05). Additional post-hoc tests would be needed to isolate the interaction.

## Example 28: Simple Blocking of $2^k$ Within-Subjects Design

**(Click in this red rectangle to see the Reference Notes on Example 28.)**

### Problem

Location in Williges (2006) Table of Contents
Section 4, Topic 17. Blocking Designs, Part 17.2.3.1. Simple Blocking Example

Page(s) in Williges (2006) Reference Material: 546-549

Problem Description
Testing was conducted on a new computerized target detection system. The detection system evaluates four different dimensions (i.e., target speed, target size, noise level, and display resolution) each with two settings. Five soldiers have been recruited to participate in the testing of the new system. For each of the 16 dimension combinations, 100 detection trials per soldier are completed and a percentage is computed. Because of the number of trials (1600 trials per soldier), the testing procedure is too lengthy to complete in one day, so it will be conducted in two sessions over two days. Do the settings have an effect on the percentage of targets detected? ($p < 0.01$) Also, is there an effect due to the blocking of the data collection into two sessions? ($p < 0.01$)

Context/Purpose
Determine if there is a significant effect due to target speed, target size, noise level, and display resolution on percent target detection while removing the potential confounding effect of experimental sessions.

Statistical Decision Criteria
A within-subjects, simple blocking design is used to control the effect of testing sessions. The simple blocking design was constructed by using the four-way interaction of display dimensions as the defining relationship to keep main effects, two-way interactions, and three-way interactions unconfounded with testing sessions.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 28: Simple Blocking of 2^k Within-Subjects Design';
data info;
input subject $ session $ speed $ size $ level $ resolution $ probability;
lines;
1     1     0     0     0     0     0.5
2     1     0     0     0     0     0.23
3     1     0     0     0     0     0.45
4     1     0     0     0     0     0.66
5     1     0     0     0     0     0.37
1     2     0     0     0     1     0.11
2     2     0     0     0     1     0.77
3     2     0     0     0     1     0.27
4     2     0     0     0     1     0.33
5     2     0     0     0     1     0.41
```

```
1     2     0     0     1     0     0.05
2     2     0     0     1     0     0.6
3     2     0     0     1     0     0.16
4     2     0     0     1     0     0.21
5     2     0     0     1     0     0.1
1     1     0     0     1     1     0.78
2     1     0     0     1     1     0.89
3     1     0     0     1     1     0.64
4     1     0     0     1     1     0.5
5     1     0     0     1     1     0.4
1     2     0     1     0     0     0.32
2     2     0     1     0     0     0.41
3     2     0     1     0     0     0.33
4     2     0     1     0     0     0.11
5     2     0     1     0     0     0.56
1     1     0     1     0     1     0.7
2     1     0     1     0     1     0.67
3     1     0     1     0     1     0.9
4     1     0     1     0     1     0.87
5     1     0     1     0     1     0.76
1     1     0     1     1     0     0.02
2     1     0     1     1     0     0.43
3     1     0     1     1     0     0.14
4     1     0     1     1     0     0.27
5     1     0     1     1     0     0.19
1     2     0     1     1     1     0.99
2     2     0     1     1     1     0.68
3     2     0     1     1     1     0.68
4     2     0     1     1     1     0.41
5     2     0     1     1     1     0.77
1     2     1     0     0     0     0.74
2     2     1     0     0     0     0.55
3     2     1     0     0     0     0.43
4     2     1     0     0     0     0.67
5     2     1     0     0     0     0.77
1     1     1     0     0     1     0.28
2     1     1     0     0     1     0.22
3     1     1     0     0     1     0.39
4     1     1     0     0     1     0.08
5     1     1     0     0     1     0.44
1     1     1     0     1     0     0.75
2     1     1     0     1     0     0.48
3     1     1     0     1     0     0.38
4     1     1     0     1     0     0.89
5     1     1     0     1     0     0.66
1     2     1     0     1     1     0.5
2     2     1     0     1     1     0.39
3     2     1     0     1     1     0.4
4     2     1     0     1     1     0.62
5     2     1     0     1     1     0.57
1     1     1     1     0     0     0.09
2     1     1     1     0     0     0.23
3     1     1     1     0     0     0.14
4     1     1     1     0     0     0.37
5     1     1     1     0     0     0.46
1     2     1     1     0     1     0.31
2     2     1     1     0     1     0.59
```

```
3     2     1     1     0     1     0.71
4     2     1     1     0     1     0.61
5     2     1     1     0     1     0.59
1     2     1     1     1     0     0.99
2     2     1     1     1     0     0.81
3     2     1     1     1     0     0.77
4     2     1     1     1     0     0.59
5     2     1     1     1     0     0.54
1     1     1     1     1     1     0.14
2     1     1     1     1     1     0.27
3     1     1     1     1     1     0.08
4     1     1     1     1     1     0.31
5     1     1     1     1     1     0.25
;
proc glm;
class subject session speed size level resolution;
model probability = subject speed speed*subject size size*subject level
level*subject resolution resolution*subject session session*subject
speed*size speed*size*subject speed*level speed*level*subject
speed*resolution speed*resolution*subject size*level size*level*subject
size*resolution size*resolution*subject level*resolution
level*resolution*subject speed*size*level speed*size*level*subject
speed*size*resolution speed*size*resolution*subject size*level*resolution
size*level*resolution*subject speed*level*resolution
speed*level*resolution*subject/ss1;
lsmeans speed size level resolution session/alpha=0.01;
test h=speed e=speed*subject;
test h=size e=size*subject;
test h=level e=level*subject;
test h=resolution e=resolution*subject;
test h=speed*size e=speed*size*subject;
test h=speed*level e=speed*level*subject;
test h=speed*resolution e=speed*resolution*subject;
test h=size*level e=size*level*subject;
test h=size*resolution e=size*resolution*subject;
test h=level*resolution e=level*resolution*subject;
test h=speed*size*level e=speed*size*level*subject;
test h=speed*size*resolution e=speed*size*resolution*subject;
test h=size*level*resolution e=size*level*resolution*subject;
test h=speed*level*resolution e=speed*level*resolution*subject;
test h=session e=session*subject;
run;
quit;
```

## SAS Output

```
Example 28: Simple Blocking of 2^k Within-Subjects Design
1

The GLM Procedure

      Class Level Information

Class           Levels    Values

subject              5    1 2 3 4 5
```

| | | |
|---|---|---|
| session | 2 | 1 2 |
| speed | 2 | 0 1 |
| size | 2 | 0 1 |
| level | 2 | 0 1 |
| resolution | 2 | 0 1 |

Number of Observations Read 80
Number of Observations Used 80

Dependent Variable: probability

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 79 | 4.91327500 | 0.06219335 | . | . |
| Error | 0 | 0.00000000 | . | | |
| **Corrected Total** | **79** | **4.91327500** | | | |

| R-Square | Coeff Var | Root MSE | probability Mean |
|---|---|---|---|
| 1.000000 | . | . | 0.471250 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| subject | 4 | 0.06723750 | 0.01680938 | . | . |
| speed | 1 | 0.00220500 | 0.00220500 | . | . |
| subject*speed | 4 | 0.16080750 | 0.04020187 | . | . |
| size | 1 | 0.00220500 | 0.00220500 | . | . |
| subject*size | 4 | 0.04513250 | 0.01128313 | . | . |
| level | 1 | 0.01012500 | 0.01012500 | . | . |
| subject*level | 4 | 0.18141250 | 0.04535312 | . | . |
| resolution | 1 | 0.10224500 | 0.10224500 | . | . |
| subject*resolution | 4 | 0.05876750 | 0.01469188 | . | . |
| session | 1 | 0.12324500 | 0.12324500 | . | . |
| subject*session | 4 | 0.10376750 | 0.02594188 | . | . |
| speed*size | 1 | 0.12324500 | 0.12324500 | . | . |
| subject*speed*size | 4 | 0.18569250 | 0.04642313 | . | . |
| speed*level | 1 | 0.08064500 | 0.08064500 | . | . |
| subject*speed*level | 4 | 0.07059250 | 0.01764813 | . | . |
| speed*resolution | 1 | 1.24500500 | 1.24500500 | . | . |
| subjec*speed*resolut | 4 | 0.08840750 | 0.02210188 | . | . |
| size*level | 1 | 0.03612500 | 0.03612500 | . | . |
| subject*size*level | 4 | 0.04103750 | 0.01025937 | . | . |
| size*resolution | 1 | 0.21840500 | 0.21840500 | . | . |
| subject*size*resolut | 4 | 0.11198250 | 0.02799563 | . | . |
| level*resolution | 1 | 0.00180500 | 0.00180500 | . | . |
| subjec*level*resolut | 4 | 0.13228250 | 0.03307063 | . | . |

| | | | | | |
|---|---|---|---|---|---|
| speed*size*level | 1 | 0.00924500 | 0.00924500 | . | . |
| subj*spee*size*level | 4 | 0.11016750 | 0.02754188 | . | . |
| speed*size*resolutio | 1 | 0.03120500 | 0.03120500 | . | . |
| subj*spee*size*resol | 4 | 0.09713250 | 0.02428313 | . | . |
| size*level*resolutio | 1 | 0.67344500 | 0.67344500 | . | . |
| subj*size*leve*resol | 4 | 0.12306750 | 0.03076687 | . | . |
| speed*level*resoluti | 1 | 0.42340500 | 0.42340500 | . | . |
| subj*spee*leve*resol | 4 | 0.25323250 | 0.06330812 | . | . |

Least Squares Means

| speed | probability LSMEAN |
|---|---|
| 0 | 0.46600000 |
| 1 | 0.47650000 |

| size | probability LSMEAN |
|---|---|
| 0 | 0.46600000 |
| 1 | 0.47650000 |

| level | probability LSMEAN |
|---|---|
| 0 | 0.46000000 |
| 1 | 0.48250000 |

| resolution | probability LSMEAN |
|---|---|
| 0 | 0.43550000 |
| 1 | 0.50700000 |

| session | probability LSMEAN |
|---|---|
| 1 | 0.43200000 |
| 2 | 0.51050000 |

Dependent Variable: probability

Tests of Hypotheses Using the Type I MS for subject*speed as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| speed | 1 | 0.00220500 | 0.00220500 | 0.05 | 0.8263 |

Tests of Hypotheses Using the Type I MS for subject*size as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| size | 1 | 0.00220500 | 0.00220500 | 0.20 | 0.6813 |

Tests of Hypotheses Using the Type I MS for subject*level as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| level | 1 | 0.01012500 | 0.01012500 | 0.22 | 0.6612 |

Tests of Hypotheses Using the Type I MS for subject*resolution as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| resolution | 1 | 0.10224500 | 0.10224500 | 6.96 | 0.0577 |

Tests of Hypotheses Using the Type I MS for subject*speed*size as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| speed*size | 1 | 0.12324500 | 0.12324500 | 2.65 | 0.1786 |

Tests of Hypotheses Using the Type I MS for subject*speed*level as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| speed*level | 1 | 0.08064500 | 0.08064500 | 4.57 | 0.0993 |

Tests of Hypotheses Using the Type I MS for subjec*speed*resolut as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **speed*resolution** | **1** | **1.24500500** | **1.24500500** | **56.33** | **0.0017** |

Dependent Variable: probability

Tests of Hypotheses Using the Type I MS for subject*size*level as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| size*level | 1 | 0.03612500 | 0.03612500 | 3.52 | 0.1338 |

Tests of Hypotheses Using the Type I MS for subject*size*resolut as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| size*resolution | 1 | 0.21840500 | 0.21840500 | 7.80 | 0.0492 |

Tests of Hypotheses Using the Type I MS for subjec*level*resolut as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| level*resolution | 1 | 0.00180500 | 0.00180500 | 0.05 | 0.8267 |

Tests of Hypotheses Using the Type I MS for subj*spee*size*level as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| speed*size*level | 1 | 0.00924500 | 0.00924500 | 0.34 | 0.5934 |

Tests of Hypotheses Using the Type I MS for subj*spee*size*resol as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| speed*size*resolutio | 1 | 0.03120500 | 0.03120500 | 1.29 | 0.3203 |

Tests of Hypotheses Using the Type I MS for subj*size*leve*resol as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **size*level*resolutio** | **1** | **0.67344500** | **0.67344500** | **21.89** | **0.0095** |

Tests of Hypotheses Using the Type I MS for subj*spee*leve*resol as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| speed*level*resoluti | 1 | 0.42340500 | 0.42340500 | 6.69 | 0.0609 |

```
Dependent Variable: probability

   Tests of Hypotheses Using the Type I MS for subject*session as an Error Term

Source                       DF       Type I SS     Mean Square    F Value    Pr > F

session                       1      0.12324500      0.12324500       4.75    0.0948
```

### Output Explanation

The p-values for the interaction of speed and resolution (0.0017) and the interaction of size, level, and resolution (0.0095) are less than the stated significance level (0.01). Therefore, both of these interactions have a significant effect on the percentage of targets detected by the soldiers. Additional post-hoc tests are needed to isolate the interaction effects. There is not a significant effect due to blocking of the 2 sessions (p = 0.0948).

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Between-Subjects | | | | |
| Subjects (S) | 4 | 0.0672 | 0.0168 | |
| Within-Subjects | | | | |
| Session (SpeedxSizexLevelxResolution) | 1 | 0.1232 | 0.1232 | 4.75 |
| Session x S (SpeedxSizexLevelxResolutionxS) | 4 | 0.1038 | 0.0259 | |
| Speed | 1 | 0.0022 | 0.0022 | 0.05 |
| Speed x S | 4 | 0.1608 | 0.0402 | |
| Size | 1 | 0.0022 | 0.0022 | 0.20 |
| Size x S | 4 | 0.0451 | 0.0113 | |
| Level | 1 | 0.0101 | 0.0101 | 0.22 |
| Level x S | 4 | 0.1814 | 0.0453 | |
| Resolution | 1 | 0.1022 | 0.1022 | 6.96 |
| Resolution x S | 4 | 0.0588 | 0.0147 | |
| Speed x Size | 1 | 0.1232 | 0.1232 | 2.65 |
| Speed x Size x S | 4 | 0.1857 | 0.0464 | |
| Speed x Level | 1 | 0.0806 | 0.0806 | 4.57 |
| Speed x Level x S | 4 | 0.0706 | 0.0176 | |
| Speed x Resolution | 1 | 1.2450 | 1.2450 | 56.33* |
| Speed x Resolution x S | 4 | 0.0884 | 0.0221 | |
| Size x Level | 1 | 0.0361 | 0.0361 | 3.52 |
| Size x Level x S | 4 | 0.0410 | 0.0103 | |
| Size x Resolution | 1 | 0.2184 | 0.2184 | 7.80 |
| Size x Resolution x S | 4 | 0.1119 | 0.0279 | |
| Level x Resolution | 1 | 0.0018 | 0.0018 | 0.05 |
| Level x Resolution x S | 4 | 0.1323 | 0.0331 | |
| Speed x Size x Level | 1 | 0.0092 | 0.0092 | 0.34 |
| Speed x Size x Level x S | 4 | 0.1102 | 0.0275 | |
| Speed x Size x Resolution | 1 | 0.0312 | 0.0312 | 1.29 |
| Speed x Size x Resolution x S | 4 | 0.0971 | 0.0243 | |
| Speed x Level x Resolution | 1 | 0.4234 | 0.4234 | 6.69 |
| Speed x Level x Resolution x S | 4 | 0.2532 | 0.0633 | |
| Size x Level x Resolution | 1 | 0.6734 | 0.6734 | 21.89* |
| Size x Level x Resolution x S | 4 | 0.1231 | 0.0308 | |
| Total | 79 | 4.9133 | | |

## Example 29: Complex Blocking of $2^k$ Within-Subjects Design

**(Click in this red rectangle to see the Reference Notes on Example 29.)**

### Problem

Location in Williges (2006) Table of Contents
Section 4, Topic 17. Blocking Designs, Part 17.2.3.2. Complex Blocking Example

Page(s) in Williges (2006) Reference Material: 550-553

Problem Description
Testing was conducted on a new computerized target detection system. The detection system evaluates four different dimensions (i.e., target speed, target size, noise level, and display resolution) each with two settings. Five soldiers have been recruited to participate in the testing of the new system. For each of the 16 dimension combinations, 100 detection trials per soldier are completed and a percentage is computed. Because of the number of trials (1600 trials per soldier), the testing procedure is too lengthy to complete in one day and is conducted in four sessions over four days. Do the settings have an effect on the percentage of targets detected? ($p < 0.01$) Also, is there an effect due to the blocking of the data collection into four sessions? ($p < 0.01$)

Context/Purpose
Determine if there is a significant effect due to target speed, target size, noise level, and display resolution unconfounded by the four testing sessions.

Statistical Decision Criteria
A within-subjects, complex blocking design should be used to accommodate the length of the testing. The four-way interaction and the speed and size interaction are used as the two defining relationships to construct the complex blocking design so as to keep main effects unconfounded with the four testing sessions.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 29: Complex Blocking of 2^k Within-Subjects Design';
data info;
input subject $ session $ speed $ size $ level $ resolution $ probability;
lines;
1     1     0     0     0     0     0.5
2     1     0     0     0     0     0.23
3     1     0     0     0     0     0.45
4     1     0     0     0     0     0.66
5     1     0     0     0     0     0.37
1     3     0     0     0     1     0.11
2     3     0     0     0     1     0.77
3     3     0     0     0     1     0.27
4     3     0     0     0     1     0.33
5     3     0     0     0     1     0.41
```

```
1    3    0    0    1    0    0.05
2    3    0    0    1    0    0.6
3    3    0    0    1    0    0.16
4    3    0    0    1    0    0.21
5    3    0    0    1    0    0.1
1    1    0    0    1    1    0.78
2    1    0    0    1    1    0.89
3    1    0    0    1    1    0.64
4    1    0    0    1    1    0.5
5    1    0    0    1    1    0.4
1    4    0    1    0    0    0.32
2    4    0    1    0    0    0.41
3    4    0    1    0    0    0.33
4    4    0    1    0    0    0.11
5    4    0    1    0    0    0.56
1    2    0    1    0    1    0.7
2    2    0    1    0    1    0.67
3    2    0    1    0    1    0.9
4    2    0    1    0    1    0.87
5    2    0    1    0    1    0.76
1    2    0    1    1    0    0.02
2    2    0    1    1    0    0.43
3    2    0    1    1    0    0.14
4    2    0    1    1    0    0.27
5    2    0    1    1    0    0.19
1    4    0    1    1    1    0.99
2    4    0    1    1    1    0.68
3    4    0    1    1    1    0.68
4    4    0    1    1    1    0.41
5    4    0    1    1    1    0.77
1    4    1    0    0    0    0.74
2    4    1    0    0    0    0.55
3    4    1    0    0    0    0.43
4    4    1    0    0    0    0.67
5    4    1    0    0    0    0.77
1    2    1    0    0    1    0.28
2    2    1    0    0    1    0.22
3    2    1    0    0    1    0.39
4    2    1    0    0    1    0.08
5    2    1    0    0    1    0.44
1    2    1    0    1    0    0.75
2    2    1    0    1    0    0.48
3    2    1    0    1    0    0.38
4    2    1    0    1    0    0.89
5    2    1    0    1    0    0.66
1    4    1    0    1    1    0.5
2    4    1    0    1    1    0.39
3    4    1    0    1    1    0.4
4    4    1    0    1    1    0.62
5    4    1    0    1    1    0.57
1    1    1    1    0    0    0.09
2    1    1    1    0    0    0.23
3    1    1    1    0    0    0.14
4    1    1    1    0    0    0.37
5    1    1    1    0    0    0.46
1    3    1    1    0    1    0.31
2    3    1    1    0    1    0.59
```

```
3     3     1     1     0     1     0.71
4     3     1     1     0     1     0.61
5     3     1     1     0     1     0.59
1     3     1     1     1     0     0.99
2     3     1     1     1     0     0.81
3     3     1     1     1     0     0.77
4     3     1     1     1     0     0.59
5     3     1     1     1     0     0.54
1     1     1     1     1     1     0.14
2     1     1     1     1     1     0.27
3     1     1     1     1     1     0.08
4     1     1     1     1     1     0.31
5     1     1     1     1     1     0.25
;
proc glm;
class subject session speed size level resolution;
model probability = subject speed speed*subject size size*subject level
level*subject resolution resolution*subject session session*subject
speed*level speed*level*subject speed*resolution speed*resolution*subject
size*level size*level*subject size*resolution size*resolution*subject
speed*size*level speed*size*level*subject speed*size*resolution
speed*size*resolution*subject size*level*resolution
size*level*resolution*subject speed*level*resolution
speed*level*resolution*subject/ss1;
lsmeans speed size level resolution/alpha=0.01;
test h=speed e=speed*subject;
test h=size e=size*subject;
test h=level e=level*subject;
test h=resolution e=resolution*subject;
test h=speed*level e=speed*level*subject;
test h=speed*resolution e=speed*resolution*subject;
test h=size*level e=size*level*subject;
test h=size*resolution e=size*resolution*subject;
test h=speed*size*level e=speed*size*level*subject;
test h=speed*size*resolution e=speed*size*resolution*subject;
test h=size*level*resolution e=size*level*resolution*subject;
test h=speed*level*resolution e=speed*level*resolution*subject;
test h=session e=session*subject;
run;
quit;
```

## SAS Output

```
Example 29: Complex Blocking of 2^k Within-Subjects Design                    1

The GLM Procedure

      Class Level Information

Class           Levels    Values

subject              5    1 2 3 4 5

session              4    1 2 3 4
```

| | | |
|---|---|---|
| speed | 2 | 0 1 |
| size | 2 | 0 1 |
| level | 2 | 0 1 |
| resolution | 2 | 0 1 |

Number of Observations Read 80
Number of Observations Used 80

Dependent Variable: probability

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 79 | 4.91327500 | 0.06219335 | . | . |
| Error | 0 | 0.00000000 | . | | |
| **Corrected Total** | **79** | **4.91327500** | | | |

| R-Square | Coeff Var | Root MSE | probability Mean |
|---|---|---|---|
| 1.000000 | . | . | 0.471250 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| subject | 4 | 0.06723750 | 0.01680938 | . | . |
| speed | 1 | 0.00220500 | 0.00220500 | . | . |
| subject*speed | 4 | 0.16080750 | 0.04020187 | . | . |
| size | 1 | 0.00220500 | 0.00220500 | . | . |
| subject*size | 4 | 0.04513250 | 0.01128313 | . | . |
| level | 1 | 0.01012500 | 0.01012500 | . | . |
| subject*level | 4 | 0.18141250 | 0.04535312 | . | . |
| resolution | 1 | 0.10224500 | 0.10224500 | . | . |
| subject*resolution | 4 | 0.05876750 | 0.01469188 | . | . |
| session | 3 | 0.24829500 | 0.08276500 | . | . |
| subject*session | 12 | 0.42174250 | 0.03514521 | . | . |
| speed*level | 1 | 0.08064500 | 0.08064500 | . | . |
| subject*speed*level | 4 | 0.07059250 | 0.01764813 | . | . |
| speed*resolution | 1 | 1.24500500 | 1.24500500 | . | . |
| subjec*speed*resolut | 4 | 0.08840750 | 0.02210188 | . | . |
| size*level | 1 | 0.03612500 | 0.03612500 | . | . |
| subject*size*level | 4 | 0.04103750 | 0.01025937 | . | . |
| size*resolution | 1 | 0.21840500 | 0.21840500 | . | . |
| subject*size*resolut | 4 | 0.11198250 | 0.02799563 | . | . |
| speed*size*level | 1 | 0.00924500 | 0.00924500 | . | . |
| subj*spee*size*level | 4 | 0.11016750 | 0.02754188 | . | . |
| speed*size*resolutio | 1 | 0.03120500 | 0.03120500 | . | . |
| subj*spee*size*resol | 4 | 0.09713250 | 0.02428312 | . | . |
| size*level*resolutio | 1 | 0.67344500 | 0.67344500 | . | . |
| subj*size*leve*resol | 4 | 0.12306750 | 0.03076688 | . | . |
| speed*level*resoluti | 1 | 0.42340500 | 0.42340500 | . | . |
| subj*spee*leve*resol | 4 | 0.25323250 | 0.06330813 | . | . |

```
Least Squares Means

          probability
speed          LSMEAN

0           0.46600000
1           0.47650000


         probability
size           LSMEAN

0          0.46600000
1          0.47650000


          probability
level          LSMEAN

0           0.46000000
1           0.48250000


               probability
resolution          LSMEAN

0               0.43550000
1               0.50700000


Dependent Variable: probability

    Tests of Hypotheses Using the Type I MS for subject*speed as an Error Term

Source                      DF       Type I SS     Mean Square    F Value    Pr > F

speed                        1      0.00220500      0.00220500       0.05    0.8263


     Tests of Hypotheses Using the Type I MS for subject*size as an Error Term

Source                      DF       Type I SS     Mean Square    F Value    Pr > F

size                         1      0.00220500      0.00220500       0.20    0.6813


    Tests of Hypotheses Using the Type I MS for subject*level as an Error Term

Source                      DF       Type I SS     Mean Square    F Value    Pr > F

level                        1      0.01012500      0.01012500       0.22    0.6612


  Tests of Hypotheses Using the Type I MS for subject*resolution as an Error Term

Source                      DF       Type I SS     Mean Square    F Value    Pr > F
```

| | | | | |
|---|---|---|---|---|
| resolution | 1 | 0.10224500 | 0.10224500 | 6.96 | 0.0577 |

Tests of Hypotheses Using the Type I MS for subject*speed*level as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| speed*level | 1 | 0.08064500 | 0.08064500 | 4.57 | 0.0993 |

Tests of Hypotheses Using the Type I MS for subjec*speed*resolut as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **speed*resolution** | **1** | **1.24500500** | **1.24500500** | **56.33** | **0.0017** |

Tests of Hypotheses Using the Type I MS for subject*size*level as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| size*level | 1 | 0.03612500 | 0.03612500 | 3.52 | 0.1338 |

Tests of Hypotheses Using the Type I MS for subject*size*resolut as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| size*resolution | 1 | 0.21840500 | 0.21840500 | 7.80 | 0.0492 |

Tests of Hypotheses Using the Type I MS for subj*spee*size*level as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| speed*size*level | 1 | 0.00924500 | 0.00924500 | 0.34 | 0.5934 |

Tests of Hypotheses Using the Type I MS for subj*spee*size*resol as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| speed*size*resolutio | 1 | 0.03120500 | 0.03120500 | 1.29 | 0.3203 |

Tests of Hypotheses Using the Type I MS for subj*size*leve*resol as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **size*level*resolutio** | **1** | **0.67344500** | **0.67344500** | **21.89** | **0.0095** |

Tests of Hypotheses Using the Type I MS for subj*spee*leve*resol as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| speed*level*resoluti | 1 | 0.42340500 | 0.42340500 | 6.69 | 0.0609 |

```
   Tests of Hypotheses Using the Type I MS for subject*session as an Error Term

Source                       DF       Type I SS     Mean Square    F Value    Pr > F

session                       3      0.24829500      0.08276500       2.35    0.1233
```

### Output Explanation

The p-values for the interaction of speed and resolution (0.0017) and the interaction of size, level, and resolution (0.0095) are less than the stated significance level (0.01). Therefore, both of these interactions have a significant effect on the percentage of targets detected by the soldiers. Note these results are similar to those in the previous example with two sessions. The number of sessions does not have a significant effect on the outcome at the 0.05 level. Additional post-hoc tests would be needed to isolate the interaction effects.

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Between-Subjects | | | | |
| Subjects (S) | 4 | 0.0672 | 0.0168 | |
| | | | | |
| Within-Subjects | | | | |
| Blocks (SpeedxSizexLevelxResolution, SpeedxSize, andLevelxResolution) | 3 | 0.2483 | 0.0828 | 2.35 |
| Blocks x S (SpeedxSizexLevelxResolutionxS, SpeedxSizexS, and LevelxResolutionxS) | 12 | 0.4217 | 0.0351 | |
| Speed | 1 | 0.0022 | 0.0022 | 0.05 |
| Speed x S | 4 | 0.1608 | 0.0402 | |
| Size | 1 | 0.0022 | 0.0022 | 0.20 |
| Size x S | 4 | 0.0451 | 0.0113 | |
| Level | 1 | 0.0101 | 0.0101 | 0.22 |
| Level x S | 4 | 0.1814 | 0.0454 | |
| Resolution | 1 | 0.1022 | 0.1022 | 6.96 |
| Resolution x S | 4 | 0.0588 | 0.0147 | |
| Speed x Level | 1 | 0.0806 | 0.0806 | 4.57 |
| Speed x Level x S | 4 | 0.0706 | 0.0176 | |
| Speed x Resolution | 1 | 1.2450 | 1.2450 | 56.33* |
| Speed x Resolution x S | 4 | 0.0884 | 0.0221 | |
| Size x Level | 1 | 0.0361 | 0.0361 | 3.52 |
| Size x Level x S | 4 | 0.0410 | 0.0103 | |
| Size x Resolution | 1 | 0.2184 | 0.2184 | 7.80 |
| Size x Resolution x S | 4 | 0.1119 | 0.0279 | |
| Speed x Size x Level | 1 | 0.0092 | 0.0092 | 0.34 |
| Speed x Size x Level x S | 4 | 0.1102 | 0.0275 | |
| Speed x Size x Resolution | 1 | 0.0312 | 0.0312 | 1.29 |
| Speed x Size x Resolution x S | 4 | 0.0971 | 0.0243 | |
| Speed x Level x Resolution | 1 | 0.4234 | 0.4234 | 6.69 |
| Speed x Level x Resolution x S | 4 | 0.2532 | 0.0633 | |
| Size x Level x Resolution | 1 | 0.6734 | 0.6734 | 21.89* |
| Size x Level x Resolution x S | 4 | 0.1231 | 0.0308 | |
| Total | 79 | 4.9133 | | |

## Example 30: One-Half Replicate of $2^4$ Between-Subjects Design

**(Click in this red rectangle to see the Reference Notes on Example 30.)**

### Problem

Location in Williges (2006) Table of Contents
Section 4, Topic 18. Blocking Designs, Part 18.1.2. Computational Considerations

Page(s) in Williges (2006) Reference Material: 574-578

Problem Description
Preliminary testing was conducted on a new computerized target detection system. Two different settings of four different factors including target speed (A), target size (B), noise level (C), and display resolution (D) were evaluated. Five different soldiers completed 100 detection trials in only one treatment combination of the four factors tested to calculate the percent of targets detected. A one-half replicate of the full factorial design was used to pretest main effects and the existence of possible two-way interactions. Do the settings of any of the four main effects of target factors and two-way interactions have a significant effect on the percent of targets detected? ($p < 0.01$)

Context/Purpose
Determine if there are significant main effects and possible two-way interactions between target speed, target size, noise level, and display resolution.

Statistical Decision Criteria
A between-subjects, Resolution IV, one-half replicate design is used to keep all the main effects unconfounded and groups of two-way interactions unconfounded from other groups. The fourth-order interaction is used as the identity relationship to construct the Resolution IV design.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 30: One-Half Replicate of 2^4 Between-Subjects Design';
data info;
input subject $ treatment $ speed $ size $ level $ resolution $ probability;
lines;
1     1     0     0     0     0     0.5
2     1     0     0     0     0     0.23
3     1     0     0     0     0     0.45
4     1     0     0     0     0     0.66
5     1     0     0     0     0     0.37
6     2     0     0     1     1     0.78
7     2     0     0     1     1     0.89
8     2     0     0     1     1     0.64
9     2     0     0     1     1     0.5
10    2     0     0     1     1     0.4
11    3     0     1     0     1     0.7
12    3     0     1     0     1     0.67
```

```
13    3     0     1     0     1     0.9
14    3     0     1     0     1     0.87
15    3     0     1     0     1     0.76
16    4     0     1     1     0     0.02
17    4     0     1     1     0     0.43
18    4     0     1     1     0     0.14
19    4     0     1     1     0     0.27
20    4     0     1     1     0     0.19
21    5     1     0     0     1     0.28
22    5     1     0     0     1     0.22
23    5     1     0     0     1     0.39
24    5     1     0     0     1     0.08
25    5     1     0     0     1     0.44
26    6     1     0     1     0     0.75
27    6     1     0     1     0     0.48
28    6     1     0     1     0     0.38
29    6     1     0     1     0     0.89
30    6     1     0     1     0     0.66
31    7     1     1     0     0     0.09
32    7     1     1     0     0     0.23
33    7     1     1     0     0     0.14
34    7     1     1     0     0     0.37
35    7     1     1     0     0     0.46
36    8     1     1     1     1     0.14
37    8     1     1     1     1     0.27
38    8     1     1     1     1     0.08
39    8     1     1     1     1     0.31
40    8     1     1     1     1     0.25
;
proc glm;
class subject treatment speed size level resolution;
model probability = speed size level speed*size speed*level size*level
resolution subject(speed*size*level*resolution)/ss1;
means speed size level resolution/alpha=0.01;
test h=speed e=subject(speed*size*level*resolution);
test h=size e=subject(speed*size*level*resolution);
test h=level e=subject(speed*size*level*resolution);
test h=speed*size e=subject(speed*size*level*resolution);
test h=speed*level e=subject(speed*size*level*resolution);
test h=size*level e=subject(speed*size*level*resolution);
test h=resolution e=subject(speed*size*level*resolution);
run;
quit;
```

## SAS Output

```
Example 30: One-Half Replicate of 2^4 Between-Subjects Design                    1

The GLM Procedure

                                Class Level Information

Class         Levels  Values

subject           40  1 10 11 12 13 14 15 16 17 18 19 2 20 21 22 23 24 25 26 27 28 29 3 30 31 32
                      33 34 35 36 37 38 39 4 40 5 6 7 8 9

treatment          8  1 2 3 4 5 6 7 8

speed              2  0 1

size               2  0 1

level              2  0 1

resolution         2  0 1


Number of Observations Read          40
Number of Observations Used          40


Dependent Variable: probability

                                   Sum of
Source                     DF         Squares     Mean Square    F Value    Pr > F

Model                      39      2.56084000      0.06566256        .       .

Error                       0      0.00000000       .

Corrected Total            39      2.56084000


R-Square     Coeff Var      Root MSE    probability Mean

1.000000           .               .            0.432000


Source                     DF       Type I SS     Mean Square    F Value    Pr > F

speed                       1      0.29929000      0.29929000        .       .
size                        1      0.18225000      0.18225000        .       .
level                       1      0.00289000      0.00289000        .       .
speed*size                  1      0.07744000      0.07744000        .       .
speed*level                 1      0.28224000      0.28224000        .       .
size*level                  1      0.85264000      0.85264000        .       .
resolution                  1      0.08649000      0.08649000        .       .
sub(spe*siz*lev*res)       32      0.77760000      0.02430000        .       .
```

| Level of speed | N | probability Mean | probability Std Dev |
|---|---|---|---|
| 0 | 20 | 0.51850000 | 0.26342431 |
| 1 | 20 | 0.34550000 | 0.22279268 |

| Level of size | N | probability Mean | probability Std Dev |
|---|---|---|---|
| 0 | 20 | 0.49950000 | 0.22382971 |
| 1 | 20 | 0.36450000 | 0.27402411 |

| Level of level | N | probability Mean | probability Std Dev |
|---|---|---|---|
| 0 | 20 | 0.44050000 | 0.25010471 |
| 1 | 20 | 0.42350000 | 0.26847082 |

| Level of resolution | N | probability Mean | probability Std Dev |
|---|---|---|---|
| 0 | 20 | 0.38550000 | 0.23098132 |
| 1 | 20 | 0.47850000 | 0.27726626 |

Dependent Variable: probability

Tests of Hypotheses Using the Type I MS for sub(spe*siz*lev*res) as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **speed** | **1** | **0.29929000** | **0.29929000** | **12.32** | **0.0014** |
| **size** | **1** | **0.18225000** | **0.18225000** | **7.50** | **0.0100** |
| level | 1 | 0.00289000 | 0.00289000 | 0.12 | 0.7325 |
| speed*size | 1 | 0.07744000 | 0.07744000 | 3.19 | 0.0837 |
| **speed*level** | **1** | **0.28224000** | **0.28224000** | **11.61** | **0.0018** |
| **size*level** | **1** | **0.85264000** | **0.85264000** | **35.09** | **<.0001** |
| resolution | 1 | 0.08649000 | 0.08649000 | 3.56 | 0.0683 |

### Output Explanation

The main effects of target speed (0.0014) and size (0.0100) are less than or equal to the stated p-value (0.01) and have a significant effect on the percentage of targets detected. The p-values for the interaction of target speed and noise level (0.0018) and the interaction of target size and noise level (<0.0001) are less than the stated significance level (0.01). The SAS printout does not show the alias structure of the fractional-factorial design. Note that the target size x display resolution is confounded with the speed x noise level interaction, and the target speed x display resolution interaction is confounded with the target size x noise level interaction in this one-half fractional factorial design. Additional data collection is needed to resolve these interactions. See Williges (2006) for a complete breakdown of the ANOVA summary table showing the alias structure.

## Example 31: 4x4 Latin Square Designs

**(Click in this red rectangle to see the Reference Notes on Example 31.)**

### Problem

Location in Williges (2006) Table of Contents
Section 4, Topic 18. Fractional-Factorial ANOVA Designs, Part 18.2.3.4. Latin Square Examples

Page(s) in Williges (2006) Reference Material: 605-609

Problem Description
The main effects of three characteristics of a hand held communication device was evaluated by forward observers in Army training exercises. Four different levels each of Input Display Color Resolution (A), Speaker Characteristics (B), and Keys Size (C), of the devices were evaluated in a 4x4 standard Latin square design. The minutes to complete a communication were measured on four soldiers in each treatment combination. Did any of the three characteristics of the communication devices have a significant effect on time to communicate ($p < 0.01$)?

Context/Purpose
Evaluate the time to complete a communication using16 configurations of a hand held communication device used by forward observers. These 16 configurations were based on four levels each of input display color resolution, speaker characteristics, and key size of the hand held communication device.

Statistical Decision Criteria
Use a 4x4 standard Latin square to test the significance of the three main effects of display color resolution, speaker characteristics, and key size ($p < 0.01$) of a hand held communication device on time to complete a communication.

### SAS Input (Part A. Between-Subjects, 4x4 Latin Square Design)

Use a between-subjects, 4x4 standard Latin square design to evaluate the hand held communication device.

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 31A: Between-Subjects, 4x4 Latin Square Design';
data six;
input Treatment Subject Color Speaker Key Time;
lines;
1 1 1 1 1 15
1 2 1 1 1 20
1 3 1 1 1 22
1 4 1 1 1 18
2 5 2 1 2 25
2 6 2 1 2 26
2 7 2 1 2 30
2 8 2 1 2 25
3 9 3 1 3 30
3 10 3 1 3 32
```

```
3 11 3 1 3 28
3 12 3 1 3 32
4 13 4 1 4 39
4 14 4 1 4 33
4 15 4 1 4 28
4 16 4 1 4 35
5 17 2 2 1 25
5 18 2 2 1 37
5 19 2 2 1 39
5 20 2 2 1 28
6 21 3 2 2 35
6 22 3 2 2 42
6 23 3 2 2 39
6 24 3 2 2 40
7 25 4 2 3 40
7 26 4 2 3 49
7 27 4 2 3 42
7 28 4 2 3 38
8 29 1 2 4 26
8 30 1 2 4 35
8 31 1 2 4 32
8 32 1 2 4 28
9 33 3 3 1 30
9 34 3 3 1 36
9 35 3 3 1 28
9 36 3 3 1 32
10 37 4 3 2 45
10 38 4 3 2 47
10 39 4 3 2 44
10 40 4 3 2 40
11 41 1 3 3 28
11 42 1 3 3 35
11 43 1 3 3 30
11 44 1 3 3 29
12 45 2 3 4 38
12 46 2 3 4 35
12 47 2 3 4 36
12 48 2 3 4 38
13 49 4 4 1 39
13 50 4 4 1 35
13 51 4 4 1 38
13 52 4 4 1 40
14 53 1 4 2 21
14 54 1 4 2 30
14 55 1 4 2 25
14 56 1 4 2 22
15 57 2 4 3 15
15 58 2 4 3 24
15 59 2 4 3 18
15 60 2 4 3 30
16 61 3 4 4 20
16 62 3 4 4 22
16 63 3 4 4 15
16 64 3 4 4 25
;
proc glm;
class Color Speaker Key;
```

```
model Time = Color Speaker Key Color*Speaker*Key;
means Color Speaker Key/alpha=0.01;
run;
quit;
```

## SAS Output (Part A. Between-Subjects, 4x4 Latin Square Design)

```
Example 31A: Between-Subjects, 4x4 Latin Square Design                                1

The GLM Procedure

    Class Level Information

Class         Levels    Values

Color              4    1 2 3 4

Speaker            4    1 2 3 4

Key                4    1 2 3 4


Number of Observations Read          64
Number of Observations Used          64


The GLM Procedure

Dependent Variable: Time

                                   Sum of
Source                     DF         Squares     Mean Square    F Value    Pr > F

Model                      15     3436.109375      229.073958      14.75    <.0001

Error                      48      745.250000       15.526042

Corrected Total            63     4181.359375


R-Square     Coeff Var      Root MSE     Time Mean

0.821768      12.59011      3.940310      31.29688


Source                     DF       Type I SS     Mean Square    F Value    Pr > F

Color                       3     1602.171875      534.057292      34.40    <.0001
Speaker                     3     1316.796875      438.932292      28.27    <.0001
Key                         3      115.171875       38.390625       2.47    0.0729
Color*Speaker*Key           6      401.968750       66.994792       4.31    0.0015


Source                     DF     Type III SS     Mean Square    F Value    Pr > F

Color                       3     1602.171875      534.057292      34.40    <.0001
Speaker                     3     1316.796875      438.932292      28.27    <.0001
```

```
Key                          3      115.171875       38.390625       2.47     0.0729
Color*Speaker*Key            6      401.968750       66.994792       4.31     0.0015


The GLM Procedure

Level of            -------------Time------------
Color          N             Mean          Std Dev

1             16       26.0000000       5.92171146
2             16       29.3125000       7.35498697
3             16       30.3750000       7.38354025
4             16       39.5000000       5.31664054


Level of            -------------Time------------
Speaker        N             Mean          Std Dev

1             16       27.3750000       6.42780419
2             16       35.9375000       6.64799469
3             16       35.6875000       6.08515954
4             16       26.1875000       8.27219237


Level of            -------------Time------------
Key            N             Mean          Std Dev

1             16       30.1250000       8.18840644
2             16       33.5000000       8.94427191
3             16       31.2500000       8.52838398
4             16       30.3125000       7.16211096
```

**Output Explanation (Part A. Between-Subjects, 4x4 Latin Square Design)**
The p-value of 0.001 for the main effects of Display Color Resolution and Type of Speaker are both less than 0.01, leading to the rejection of the null hypothesis. The locus of these two main effects requires additional post hoc analyses. In addition, the p = .002 value for Residual (i.e., Color*Speaker*Key) is significant at the specified 0.01 level. Consequently, Residual cannot be combined with Error to provide a pooled error term.

**SAS Input (Part B. Within-Subjects, 4x4 Latin Square Design)**

Use a within-subjects, 4x4 standard Latin square design to evaluate the hand held communication device.

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 31B: Within-Subjects, 4x4 Latin Square Design';
data six;
input Treatment Subject Color Speaker Key Time;
lines;
1 1 1 1 1 15
1 2 1 1 1 20
```

```
1 3 1 1 1 22
1 4 1 1 1 18
2 1 2 1 2 25
2 2 2 1 2 26
2 3 2 1 2 30
2 4 2 1 2 25
3 1 3 1 3 30
3 2 3 1 3 32
3 3 3 1 3 28
3 4 3 1 3 32
4 1 4 1 4 39
4 2 4 1 4 33
4 3 4 1 4 28
4 4 4 1 4 35
5 1 2 2 1 25
5 2 2 2 1 37
5 3 2 2 1 39
5 4 2 2 1 28
6 1 3 2 2 35
6 2 3 2 2 42
6 3 3 2 2 39
6 4 3 2 2 40
7 1 4 2 3 40
7 2 4 2 3 49
7 3 4 2 3 42
7 4 4 2 3 38
8 1 1 2 4 26
8 2 1 2 4 35
8 3 1 2 4 32
8 4 1 2 4 28
9 1 3 3 1 30
9 2 3 3 1 36
9 3 3 3 1 28
9 4 3 3 1 32
10 1 4 3 2 45
10 2 4 3 2 47
10 3 4 3 2 44
10 4 4 3 2 40
11 1 1 3 3 28
11 2 1 3 3 35
11 3 1 3 3 30
11 4 1 3 3 29
12 1 2 3 4 38
12 2 2 3 4 35
12 3 2 3 4 36
12 4 2 3 4 38
13 1 4 4 1 39
13 2 4 4 1 35
13 3 4 4 1 38
13 4 4 4 1 40
14 1 1 4 2 21
14 2 1 4 2 30
14 3 1 4 2 25
14 4 1 4 2 22
15 1 2 4 3 15
15 2 2 4 3 24
15 3 2 4 3 18
```

```
15 4 2 4 3 30
16 1 3 4 4 20
16 2 3 4 4 22
16 3 3 4 4 15
16 4 3 4 4 25
;
proc glm;
class Subject Treatment;
model Time = Subject Treatment Treatment*Subject;
test h=Treatment e=Treatment*Subject;
proc glm;
class Subject Color Speaker Key;
model Time = Subject Color Speaker Key Color*Speaker*Key Subject*Color
Subject*Speaker Subject*Key Subject*Color*Speaker*Key;
means Color Speaker Key/alpha=0.01;
test h=Color e=Color*Subject;
test h=Speaker e=Speaker*Subject;
test h=Key e=Key*Subject;
test h=Color*Speaker*Key e=Color*Speaker*Key*Subject;
run;
quit;
```

### SAS Output (Part B. Within-Subjects, 4x4 Latin Square Design)

```
Example 31B: Within-Subjects, 4x4 Latin Square Design                                 1

The GLM Procedure

                  Class Level Information

Class          Levels    Values

Subject             4    1 2 3 4

Treatment          16    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16


Number of Observations Read          64
Number of Observations Used          64


The GLM Procedure

Dependent Variable: Time

                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F

Model                       63     4181.359375       66.370784        .       .

Error                        0        0.000000         .

Corrected Total             63     4181.359375


R-Square     Coeff Var      Root MSE     Time Mean
```

```
1.000000          .              .      31.29688


Source                      DF       Type I SS     Mean Square    F Value    Pr > F

Subject                      3      144.921875       48.307292          .       .
Treatment                   15     3436.109375      229.073958          .       .
Subject*Treatment           45      600.328125       13.340625          .       .


Source                      DF     Type III SS     Mean Square    F Value    Pr > F

Subject                      3      144.921875       48.307292          .       .
Treatment                   15     3436.109375      229.073958          .       .
Subject*Treatment           45      600.328125       13.340625          .       .


 Tests of Hypotheses Using the Type III MS for Subject*Treatment as an Error Term

Source                      DF     Type III SS     Mean Square    F Value    Pr > F

Treatment                   15     3436.109375      229.073958      17.17    <.0001


The GLM Procedure

    Class Level Information

Class         Levels    Values

Subject            4    1 2 3 4

Color              4    1 2 3 4

Speaker            4    1 2 3 4

Key                4    1 2 3 4


Number of Observations Read          64
Number of Observations Used          64


The GLM Procedure

Dependent Variable: Time

                                   Sum of
Source                      DF        Squares     Mean Square    F Value    Pr > F

Model                       63     4181.359375       66.370784          .       .

Error                        0        0.000000                .

Corrected Total             63     4181.359375
```

| R-Square | Coeff Var | Root MSE | Time Mean |
|---|---|---|---|
| 1.000000 | . | . | 31.29688 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Subject | 3 | 144.921875 | 48.307292 | . | . |
| Color | 3 | 1602.171875 | 534.057292 | . | . |
| Speaker | 3 | 1316.796875 | 438.932292 | . | . |
| Key | 3 | 115.171875 | 38.390625 | . | . |
| Color*Speaker*Key | 6 | 401.968750 | 66.994792 | . | . |
| Subject*Color | 9 | 170.515625 | 18.946181 | . | . |
| Subject*Speaker | 9 | 194.890625 | 21.654514 | . | . |
| Subject*Key | 9 | 121.515625 | 13.501736 | . | . |
| Subj*Color*Speak*Key | 18 | 113.406250 | 6.300347 | . | . |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Subject | 3 | 144.921875 | 48.307292 | . | . |
| Color | 3 | 1602.171875 | 534.057292 | . | . |
| Speaker | 3 | 1316.796875 | 438.932292 | . | . |
| Key | 3 | 115.171875 | 38.390625 | . | . |
| Color*Speaker*Key | 6 | 401.968750 | 66.994792 | . | . |
| Subject*Color | 9 | **170.515625** | **18.946181** | . | . |
| Subject*Speaker | 9 | **194.890625** | **21.654514** | . | . |
| Subject*Key | 9 | **121.515625** | **13.501736** | . | . |
| Subj*Color*Speak*Key | 18 | **113.406250** | **6.300347** | . | . |

The GLM Procedure

| Level of Color | N | Time Mean | Time Std Dev |
|---|---|---|---|
| 1 | 16 | 26.0000000 | 5.92171146 |
| 2 | 16 | 29.3125000 | 7.35498697 |
| 3 | 16 | 30.3750000 | 7.38354025 |
| 4 | 16 | 39.5000000 | 5.31664054 |

| Level of Speaker | N | Time Mean | Time Std Dev |
|---|---|---|---|
| 1 | 16 | 27.3750000 | 6.42780419 |
| 2 | 16 | 35.9375000 | 6.64799469 |
| 3 | 16 | 35.6875000 | 6.08515954 |
| 4 | 16 | 26.1875000 | 8.27219237 |

| Level of Key | N | Time Mean | Time Std Dev |
|---|---|---|---|
| 1 | 16 | 30.1250000 | 8.18840644 |
| 2 | 16 | 33.5000000 | 8.94427191 |

```
3            16       31.2500000       8.52838398
4            16       30.3125000       7.16211096

The GLM Procedure

Dependent Variable: Time

   Tests of Hypotheses Using the Type III MS for Subject*Color as an Error Term

Source                       DF     Type III SS     Mean Square    F Value    Pr > F

Color                         3     1602.171875      534.057292      28.19    <.0001


  Tests of Hypotheses Using the Type III MS for Subject*Speaker as an Error Term

Source                       DF     Type III SS     Mean Square    F Value    Pr > F

Speaker                       3     1316.796875      438.932292      20.27    0.0002


    Tests of Hypotheses Using the Type III MS for Subject*Key as an Error Term

Source                       DF     Type III SS     Mean Square    F Value    Pr > F

Key                           3     115.1718750      38.3906250       2.84    0.0979


Tests of Hypotheses Using the Type III MS for Subj*Color*Speak*Key as an Error Term

Source                       DF     Type III SS     Mean Square    F Value    Pr > F

Color*Speaker*Key             6     401.9687500      66.9947917      10.63    <.0001
```

### Output Explanation (Part B. Within-Subjects, 4x4 Latin Square Design)

The p-value of less than 0.001 for the main effect of Display Color Resolution and the p-value of 0.002 for the main effect of Type of Speaker are both less than 0.01, leading to the rejection of the null hypothesis. The locus of these two main effects requires additional post hoc analyses. In addition, the $p < 0.0001$ value for Residual (i.e., Color*Speaker*Key) is significant at the specified 0.01 level. Consequently, Residual cannot be combined with Subject*Color, Subjects*Speaker, Subject*Key, and Subject*Color*Speaker*Key to provide a pooled error term.

## Example 32: Linear Correlation Coefficient

**(Click in this red rectangle to see the Reference Notes on Example 32.)**

### Problem

Location in Williges (2006) Table of Contents
Section 4, Topic 19. Analysis of Covariance (ANCOVA), Part 19.2.1.1. Computational Formulae

Page(s) in Williges (2006) Reference Material: 623-633

Problem Description
The Army is trying to update their anthropometric database. They are currently recording the height, weight, age, and gender of new recruits that are enlisting. First they would like to determine the degree of linear relationship of height and weight and if this relationship is significant ($p < 0.05$).

Context/Purpose
Determine the degree of the linear relationship between the factors to ensure that the measurements are valid.

Statistical Decision Criteria
Calculate the Pearson Product Moment correlation and test for significance.

### SAS Input (Part A. Pearson Product Moment Correlation and Significance Test)

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 32A: Pearson Product Moment Correlation and Significance Test';
data info;
input height weight;
lines;
68 190
62 133
71 132
76 211
72 200
67 154
63 125
75 158
78 179
65 139
70 188
69 191
70 155
69 140
64 120
70 188
;
proc corr pearson;
run;
quit;
```

### SAS Output (Part A. Pearson Product Moment Correlation and Significance Test)

```
Example 32A: Pearson Product Moment Correlation and Significance Test                    1

The CORR Procedure

   2  Variables:    height   weight


                                  Simple Statistics

Variable           N          Mean       Std Dev           Sum       Minimum       Maximum

height            16      69.31250       4.55659          1109      62.00000      78.00000
weight            16     162.68750      29.57413          2603     120.00000     211.00000


Pearson Correlation Coefficients, N = 16
       Prob > |r| under H0: Rho=0

              height        weight

height       1.00000       0.63451
                           0.0083

weight       0.63451       1.00000
              0.0083
```

### Output Explanation (Part A. Pearson Product-Moment Correlation and Significance Test)

The correlation value (r = 0.635) indicates that the relationship between the two factors, height and weight are positively correlated. The p-value (0.0083) is less than the stated value (0.05), indicating that it is statistically significant.

### SAS Input (Part B. Comparison of Two Correlations Significance Test)

The Army would like to further investigate the degree of the relationship between height and weight comparing each gender (Female = 1 and Male = 2). Is there a significant difference ($p < 0.05$) between the correlations of height and weight for six female ($r_1 = 0.648$) and six male ($r_2 = 0.615$) soldiers?

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 32B: Comparison of Two Correlations Significance Test';
data info;
input height weight gender;
lines;
68 190 1
62 133 1
71 132 1
76 211 1
72 200 1
67 154 1
```

```
63 125 1
75 158 1
78 179 2
65 139 2
70 188 2
69 191 2
70 155 2
69 140 2
64 120 2
70 188 2
;
proc corr data=info (where=(gender=1 or gender=2)) fisher;
var height weight;
by gender;
run;
quit;
```

## SAS Output (Part B. Comparison of Two Correlations Significance Test)

```
Example 32B: Comparison of Two Correlations Significance Test                    1

gender=1

The CORR Procedure

   2  Variables:    height   weight

                                Simple Statistics

Variable           N          Mean       Std Dev           Sum       Minimum       Maximum

height             8      69.25000       5.17549     554.00000      62.00000      76.00000
weight             8     162.87500      33.40846          1303     125.00000     211.00000


Pearson Correlation Coefficients, N = 8
       Prob > |r| under HO: Rho=0

              height        weight

height       1.00000       0.64796
                           0.0823

weight       0.64796       1.00000
              0.0823

               Pearson Correlation Statistics (Fisher's z Transformation)

           With                          Sample                       Bias    Correlation
Variable   Variable           N      Correlation     Fisher's z    Adjustment       Estimate

height     weight             8          0.64796        0.77177       0.04628        0.62030
```

```
Pearson Correlation Statistics (Fisher's z Transformation)

            With                                    p Value for
Variable    Variable       95% Confidence Limits       HO:Rho=0

height      weight         -0.149892      0.921971        0.0844


gender=2

The CORR Procedure

   2  Variables:    height   weight

                                  Simple Statistics

Variable           N          Mean       Std Dev           Sum       Minimum       Maximum

height             8      69.37500       4.20671     555.00000      64.00000      78.00000
weight             8     162.50000      27.53180          1300     120.00000     191.00000

Pearson Correlation Coefficients, N = 8
       Prob > |r| under HO: Rho=0

              height        weight

height       1.00000       0.61488
                           0.1047

weight       0.61488       1.00000
             0.1047

               Pearson Correlation Statistics (Fisher's z Transformation)

            With                         Sample                       Bias    Correlation
Variable    Variable           N     Correlation    Fisher's z    Adjustment       Estimate

height      weight             8         0.61488       0.71673       0.04392        0.58682

   Pearson Correlation Statistics (Fisher's z Transformation)

            With                                    p Value for
Variable    Variable       95% Confidence Limits       HO:Rho=0

height      weight         -0.200942      0.913675        0.1090
```

### Output Explanation (Part B. Comparison of Two Correlations Significance Test)

The correlation value (r = 0.648) indicates that the two factors, height and weight, are positively correlated when gender is female (gender=1). The p-value (0.082) is greater than the stated value (0.05) indicating that it is not statistically significant. The correlation value (r = 0.615) indicates that the two factors, height and weight are positively correlated when gender is male (gender=2). The p-value (0.105) is greater than the stated value (0.05) indicating that it is not statistically significant. SAS does not calculate the significance test of the two correlations. The hand calculations can be found in the Williges (2006) reference.

## Example 33: Alternative Linear Correlations

**(Click in this red rectangle to see the Reference Notes on Example 33.)**

### Problem

Location in Williges (2006) Table of Contents
Section 4, Topic 19. Analysis of Covariance (ANCOVA), Part 19.2.2. Alternative Correlations

Page(s) in Williges (2006) Reference Material: 636-653

Problem Description
A study is conducted to determine if there is a relationship between the number of years of service for sixteen officers and their current enlistment status (1=enlisted, 0=officer). The gender (1=male, 0=female) of the officers and the expected duration (in years) at their current post was also recorded. Before defining the full relationship between these factors, the researchers want to determine the correlation of these variables. Nonparametric correlations must be used because the variables are either dichotomous or rank ordered.

Context/Purpose
Determine the degree of the linear relationship between the factors of interest.

Statistical Decision Criteria
Conduct various nonparametric correlations between the classification and rank ordered variables to assess the linear relationships among them and conduct partial and part correlations to account for the effect of a third variable on the correlation.

### SAS Input (Part A. Point Biserial Correlation)

First, the researchers would like to know the degree of the relationship of the officers' enlistment status and the number of years that they have been in the Army which requires a point biserial correlation since one variable is continuous and one is dichotomous.

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 33A: Point Biserial Correlation';
data info;
input status years;
lines;
1     20
1     24
1     27
0     29
1     30
1     31
1     33
1     34
0     35
1     36
0     37
```

```
0     38
0     39
0     40
0     42
0     44
;
proc corr pearson;
run;
quit;
```

### SAS Output (Part A. Point Biserial)

```
Example 33A: Point Biserial Correlation                                    1

The CORR Procedure

   2  Variables:    status   years


                                Simple Statistics

Variable           N          Mean       Std Dev           Sum       Minimum       Maximum

status            16       0.50000       0.51640       8.00000             0       1.00000
years             16      33.68750       6.57996     539.00000      20.00000      44.00000


Pearson Correlation Coefficients, N = 16
       Prob > |r| under H0: Rho=0

              status         years

status       1.00000      -0.67689
                           0.0040

years       -0.67689       1.00000
             0.0040
```

### Output Explanation (Part A. Point Biserial Correlation)

For SAS to output the point biserial correlation, a Pearson correlation is conducted where one of the variables is dichotomous. The point biserial correlation value (r = -0.677) indicates that the variables are negatively correlated. The p-value of the F-test (0.0040) is less than the stated value (0.05) indicating that this correlation is statistically significant.

### SAS Input (Part B. Phi Correlation)

Next, the researchers would like to determine the degree of the relationship between the officers' enlistment status and their gender. Determining the degree of relationship requires a phi correlation since both variables are dichotomous.

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 33B: Phi Correlation';
```

```
data info;
input status gender;
lines;
1 1
0 0
0 0
1 1
0 0
1 0
1 0
1 1
0 1
0 0
1 1
0 1
1 0
1 0
1 1
0 0
;
proc corr pearson;
run;
quit;
```

### SAS Output (Part B. Phi Correlation)

```
Example 33B: Phi Correlation                                                    1

The CORR Procedure

   2  Variables:    status   gender


                                Simple Statistics

Variable           N          Mean       Std Dev           Sum       Minimum       Maximum

status            16       0.56250       0.51235       9.00000             0       1.00000
gender            16       0.43750       0.51235       7.00000             0       1.00000


Pearson Correlation Coefficients, N = 16
       Prob > |r| under HO: Rho=O

              status        gender

status       1.00000       0.26984
                           0.3122

gender       0.26984       1.00000
              0.3122
```

### Output Explanation (Part B. Phi Correlation)

For SAS to output the Phi correlation, a Pearson correlation is conducted where two of the variables are dichotomous. The squared Phi correlation value (r = 0.269) indicates that the variables are positively correlated. The p-value of the F-test (0.3122) is greater than the stated value (0.05) indicating that the relationship is not statistically significant.

### SAS Input (Part C. Spearman Rank-Order, Rho, Correlation)

The number of years of service and the remaining number of months the officers believe they will be stationed at their post were converted into rank orders. What is the Spearman rank-order correlation between these two rank orders and is this correlation significant ($p < 0.05$)?

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 33C: Spearman Rho Correlation';
data info;
input years remaining;
lines;
1     4
2     5
3     6
4     2
5     3
6     10
7     9
8     1
9     14
10    13
11    15
12    16
13    7
14    11
15    12
16    8
;
proc corr data=info spearman;
var years remaining;
run;
quit;
```

### SAS Output (Part C. Spearman Rank-Order, Rho, Correlation)

```
Example 33C: Spearman Rho Correlation                                    1

The CORR Procedure

   2  Variables:    years     remaining


                                Simple Statistics

Variable            N          Mean       Std Dev        Median       Minimum       Maximum

years              16       8.50000       4.76095       8.50000       1.00000      16.00000
```

```
remaining          16       8.50000       4.76095       8.50000       1.00000       16.00000


Spearman Correlation Coefficients, N = 16
        Prob > |r| under HO: Rho=0

                  years      remaining

years           1.00000        0.57647
                                0.0194

remaining       0.57647        1.00000
                 0.0194
```

**Output Explanation (Part C. Spearman Rank-Order, Rho, Correlation)**
The squared Spearman correlation value (r = 0.576) indicates that the variables, years of service and time remaining at their current post, are positively correlated. The p-value of the F-test (0.02) is less than the stated value (0.05) indicating that it is statistically significant.

**SAS Input (Part D. Partial Correlation)**

Next the Army would like to determine the degree of the relationship between height and weight when the factor of age is held constant by using a partial correlation.

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 33D: Partial Correlation';
data info;
input height weight age;
lines;
68 190 22
62 133 19
71 132 18
76 211 22
72 200 26
67 154 19
63 125 22
75 158 25
78 179 19
65 139 18
70 188 25
69 191 18
70 155 23
69 140 23
64 120 20
70 188 21
;
proc corr pearson;
run;
proc corr data=info;
var height weight;
partial age;
```

```
run;
quit;
```

## SAS Output (Part D. Partial Correlation)

```
Example 33D: Partial Correlation                                                1

The CORR Procedure

   3  Variables:    height   weight   age


                                Simple Statistics

Variable           N          Mean       Std Dev           Sum       Minimum       Maximum

height            16      69.31250       4.55659          1109      62.00000      78.00000
weight            16     162.68750      29.57413          2603     120.00000     211.00000
age               16      21.25000       2.67083     340.00000      18.00000      26.00000


   Pearson Correlation Coefficients, N = 16
           Prob > |r| under H0: Rho=0

             height        weight           age

height       1.00000       0.63451       0.29992
                            0.0083        0.2591

weight       0.63451       1.00000       0.34710
             0.0083                       0.1878

age          0.29992       0.34710       1.00000
             0.2591        0.1878


The CORR Procedure

   1 Partial Variables:    age
   2         Variables:    height   weight

                                Simple Statistics

Variable           N          Mean       Std Dev           Sum       Minimum       Maximum

age               16      21.25000       2.67083     340.00000      18.00000      26.00000
height            16      69.31250       4.55659          1109      62.00000      78.00000
weight            16     162.68750      29.57413          2603     120.00000     211.00000

         Simple Statistics

              Partial       Partial
Variable     Variance       Std Dev

age
height        20.24449       4.49939
weight       824.20110      28.70890
```

```
Pearson Partial Correlation Coefficients, N = 16
       Prob > |r| under HO: Partial Rho=O

              height        weight

height       1.00000       0.59286
                            0.0198

weight       0.59286       1.00000
              0.0198
```

**Output Explanation (Part D. Partial Correlation)**
The correlation value (r = 0.593) indicates that the relationship between height and weight when age is held constant is positively correlated. The p-value (0.02) is less than the stated value (0.05) indicating that it is statistically significant.

**SAS Input (Part E. Semi-Partial Correlation)**

The Army has studied the correlation between height and weight when age is removed in the previous examples, but they would now like to know the correlation between height and weight when only the variance in common between weight and age are removed by calculating a semi-partial correlation.

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 33E: Semi-Partial Correlation';
data info;
input height weight age;
lines;
68 190 22
62 133 19
71 132 18
76 211 22
72 200 26
67 154 19
63 125 22
75 158 25
78 179 19
65 139 18
70 188 25
69 191 18
70 155 23
69 140 23
64 120 20
70 188 21
;
proc reg data=info;
model weight= height age/pcorr2 scorr2(tests);
run;
quit;
```

## SAS Output (Part E. Semi-Partial Correlation)

```
Example 33E: Semi-Partial Correlation                                    1

The REG Procedure
Model: MODEL1
Dependent Variable: weight

Number of Observations Read          16
Number of Observations Used          16


                          Analysis of Variance

                                   Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F

Model                       2    5636.29141    2818.14570       4.90    0.0260
Error                      13    7483.14609     575.62662
Corrected Total            15         13119


Root MSE              23.99222    R-Square     0.4296
Dependent Mean       162.68750    Adj R-Sq     0.3419
Coeff Var             14.74743


                             Parameter Estimates

                                                                  Squared
                  Parameter     Standard                      Semi-partial  ----Type II----
Variable   DF      Estimate        Error  t Value  Pr > |t|  Corr Type II  F Value  Pr > F

Intercept   1   -140.05020     96.96051    -1.44    0.1723            .       .     .
height      1      3.78280      1.42513     2.65    0.0198      0.30913    7.05  0.0198
age         1      1.90786      2.43134     0.78    0.4467      0.02702    0.62  0.4467

    Parameter Estimates

                    Squared
                    Partial
Variable   DF  Corr Type II

Intercept   1             .
height      1       0.35148
age         1       0.04522
```

## Output Explanation (Part E. Semi-Partial Correlation)

The semi-partial (r = 0.556) indicates there is positive relationship between height and weight when the correlation between weight and age is removed. The p-value of the F-test (0.02) on the squared semi-partial correlation value ($r^2$ = 0.309) is less than the stated value (0.05) indicating that it is statistically significant.

## Example 34: Simple Linear Regression

**(Click in this red rectangle to see the Reference Notes on Example 34.)**

### Problem

Location in Williges (2006) Table of Contents
Section 4, Topic 19. Analysis of Covariance (ANCOVA), Part 19.3.1.2. Calculation Example

Page(s) in Williges (2006) Reference Material: 659-660, 666-669

Problem Description
The Army is currently recording the height (X) and weight (Y) of new recruits that are enlisting. To what extent can weight of Army recruits be predicted by their height and is this prediction significant ($p < 0.01$)?

Context/Purpose
Determine the extent to which Army recruit weight can be predicted by height.

Statistical Decision Criteria
Conduct a simple linear regression to predict weight as a function of height and test the significance of the prediction at the 0.01 level of significance.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 34: Simple Linear Regression';
data info;
input height weight;
lines;
68 190
62 133
71 132
76 211
72 200
67 154
63 125
75 158
78 179
65 139
70 188
69 191
70 155
69 140
64 120
70 188
;
proc glm data=info;
model weight=height/alpha=0.05 P;
output out=prediction
```

```
      p = predweight;
run;
proc corr;
var weight predweight;
run;
quit;
```

## SAS Output

```
Example 34: Simple Linear Regression                                    1

The GLM Procedure

Number of Observations Read          16
Number of Observations Used          16

Dependent Variable: weight

                                   Sum of
Source                     DF         Squares     Mean Square    F Value    Pr > F

Model                       1      5281.85131      5281.85131       9.43    0.0083

Error                      14      7837.58619       559.82759

Corrected Total            15     13119.43750



R-Square     Coeff Var      Root MSE    weight Mean

0.402597      14.54363      23.66068       162.6875


Source                     DF       Type I SS     Mean Square    F Value    Pr > F

height                      1     5281.851307     5281.851307       9.43    0.0083


Source                     DF     Type III SS     Mean Square    F Value    Pr > F

height                      1     5281.851307     5281.851307       9.43    0.0083


                                  Standard
Parameter         Estimate           Error    t Value    Pr > |t|

Intercept     -122.7553683     93.11749347      -1.32      0.2086
height           4.1182019      1.34073114       3.07      0.0083
```

| Observation | Observed | Predicted | Residual |
|---|---|---|---|
| 1 | 190.00000000 | 157.28236002 | 32.71763998 |
| 2 | 133.00000000 | 132.57314871 | 0.42685129 |
| 3 | 132.00000000 | 169.63696568 | -37.63696568 |
| 4 | 211.00000000 | 190.22797512 | 20.77202488 |
| 5 | 200.00000000 | 173.75516757 | 26.24483243 |
| 6 | 154.00000000 | 153.16415814 | 0.83584186 |
| 7 | 125.00000000 | 136.69135059 | -11.69135059 |
| 8 | 158.00000000 | 186.10977323 | -28.10977323 |
| 9 | 179.00000000 | 198.46437889 | -19.46437889 |
| 10 | 139.00000000 | 144.92775436 | -5.92775436 |
| 11 | 188.00000000 | 165.51876380 | 22.48123620 |
| 12 | 191.00000000 | 161.40056191 | 29.59943809 |
| 13 | 155.00000000 | 165.51876380 | -10.51876380 |
| 14 | 140.00000000 | 161.40056191 | -21.40056191 |
| 15 | 120.00000000 | 140.80955248 | -20.80955248 |
| 16 | 188.00000000 | 165.51876380 | 22.48123620 |

| | |
|---|---|
| Sum of Residuals | -0.000000 |
| Sum of Squared Residuals | 7837.586193 |
| Sum of Squared Residuals - Error SS | 0.000000 |
| First Order Autocorrelation | 0.151581 |
| Durbin-Watson D | 1.495776 |

The CORR Procedure

2 Variables: weight predweight

Simple Statistics

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| weight | 16 | 162.68750 | 29.57413 | 2603 | 120.00000 | 211.00000 |
| predweight | 16 | 162.68750 | 18.76495 | 2603 | 132.57315 | 198.46438 |

Pearson Correlation Coefficients, N = 16
Prob > |r| under HO: Rho=0

| | weight | predweight |
|---|---|---|
| weight | 1.00000 | **0.63451**<br>**0.0083** |
| predweight | **0.63451**<br>**0.0083** | 1.00000 |

**Output Explanation**

The least squares solution for predicted weight as a function of height is expressed in a simple linear regression equation (Weight = -122.76 + 4.12Height). The results of the ANOVA performed on this simple regression show that the partial regression of weight on height is significant at the p = 0.0083 level, which is less than the stated value (0.05). Since simple regression uses only one predictor, the p-value of the simple linear regression and the correlation between height and weight in this example match the p-value and the Pearson correlation in Example 31A which used the same data set.

## Example 35: One-Way, Analysis of Covariance (ANCOVA)

**(Click in this red rectangle to see the Reference Notes on Example 35.)**

### Problem

Location in Williges (2006) Table of Contents
Section 4, Topic 19. Analysis of Covariance (ANCOVA), Part 19.4.1. Basic ANCOVA Design

Page(s) in Williges (2006) Reference Material: 673-682

Problem Description
An experiment was conducted to study the effects of three different weight training methods used during basic training. One group of eight soldiers used basic weight training (A), another group of eight soldiers received weight training and aerobic exercise (B), and a third group of eight soldiers received weight training and diet control (C). The maximum lifting weight (MLW) of the 24 soldiers was measured after two months of training on one of the three methods. A covariate, the weight of each subject was measured before measurement of MLW. Were the three different weight training methods significantly different ($p < 0.05$) in terms of MLW?

Context/Purpose
Determine the differences among three weight lifting training programs in terms of MLW after two months of training.

Statistical Decision Criteria
Conduct a between-subjects ANOVA on the three training programs and conduct an ANCOVA on the three training programs using weight of soldier as the covariate.

### SAS Input (Part A: ANOVA One-Way, Between-Subjects)

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 35A: ANOVA One-Way, Between-Subjects Design';
data info;
input  group $ weight MLW;
lines;
A     183   240
A     168   264
A     220   300
A     200   342
A     192   249
A     178   277
A     185   285
A     190   263
B     200   360
B     207   295
B     172   260
B     188   305
B     201   340
B     177   285
B     171   290
```

```
B     167   255
C     182   275
C     194   307
C     179   240
C     213   333
C     194   248
C     185   232
C     183   267
C     193   289
;
proc glm data=info;
class group weight;
model MLW = group;
means group;
run;
quit;
```

### SAS Output (Part A: ANOVA One-Way, Between-Subjects)

```
Example 35A: ANOVA One-Way, Between-Subjects Design                                1

The GLM Procedure

                                Class Level Information

Class       Levels  Values

group            3  A B C

weight          20  167 168 171 172 177 178 179 182 183 185 188 190 192 193 194 200 201 207 213
                    220


Number of Observations Read          24
Number of Observations Used          24

Dependent Variable: MLW

                                   Sum of
Source                     DF         Squares     Mean Square    F Value    Pr > F

Model                       2      2889.25000      1444.62500       1.22    0.3166

Error                      21     24962.37500      1188.68452

Corrected Total            23     27851.62500


R-Square     Coeff Var      Root MSE      MLW Mean

0.103737      12.16667      34.47730      283.3750
```

```
Source                     DF       Type I SS     Mean Square    F Value    Pr > F

group                       2     2889.250000     1444.625000       1.22    0.3166


Source                     DF     Type III SS     Mean Square    F Value    Pr > F

group                       2     2889.250000     1444.625000       1.22    0.3166


Level of           -------------MLW-------------
group        N             Mean          Std Dev

A            8       277.500000       32.3684149
B            8       298.750000       36.2284419
C            8       273.875000       34.7251967
```

**Output Explanation (Part A: ANOVA One-Way, Between-Subjects)**
Based on the ANOVA results there is no significant difference among the three weight training programs since the p-value (0.317) is greater than the stated p-value (0.05).

**SAS Input (Part B: Regression ANOVA One-Way, Between-Subjects)**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 35B: Regression ANOVA One-Way, Between-Subjects Design';
data info;
input  group $ weight MLW;
lines;
A     183   240
A     168   264
A     220   300
A     200   342
A     192   249
A     178   277
A     185   285
A     190   263
B     200   360
B     207   295
B     172   260
B     188   305
B     201   340
B     177   285
B     171   290
B     167   255
C     182   275
C     194   307
C     179   240
C     213   333
C     194   248
C     185   232
C     183   267
```

```
C     193   289
;
proc glm data=info;
model MLW = weight;
run;
quit;
```

### SAS Output (Part B: Regression ANOVA One-Way, Between-Subjects)

```
Example 35B: Regression ANOVA One-Way, Between-Subjects Design                          1

The GLM Procedure

Number of Observations Read          24
Number of Observations Used          24

Dependent Variable: MLW

                                   Sum of
Source                     DF         Squares     Mean Square    F Value    Pr > F

Model                       1      9111.02357      9111.02357      10.70    0.0035

Error                      22     18740.60143       851.84552

Corrected Total            23     27851.62500


R-Square     Coeff Var      Root MSE      MLW Mean

0.327127      10.29957      29.18639      283.3750


Source                     DF       Type I SS     Mean Square    F Value    Pr > F

weight                      1     9111.023574     9111.023574      10.70    0.0035


Source                     DF     Type III SS     Mean Square    F Value    Pr > F

weight                      1     9111.023574     9111.023574      10.70    0.0035


                                  Standard
Parameter         Estimate           Error    t Value    Pr > |t|

Intercept      11.37362281     83.38334306       0.14      0.8927
weight          1.44361633      0.44141656       3.27      0.0035
```

### Output Explanation (Part B: Regression ANOVA One-Way, Between-Subjects)

The partial regression value for soldier weight is significant because the p-value (0.0035) is smaller than the stated p-value (0.05). Consequently, weight is a significant covariate.

### SAS Input (Part C: ANCOVA One-Way, Between-Subjects)

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 35C: ANCOVA One-Way, Between-Subjects Design';
data info;
input  group $ weight MLW;
lines;
A     183   240
A     168   264
A     220   300
A     200   342
A     192   249
A     178   277
A     185   285
A     190   263
B     200   360
B     207   295
B     172   260
B     188   305
B     201   340
B     177   285
B     171   290
B     167   255
C     182   275
C     194   307
C     179   240
C     213   333
C     194   248
C     185   232
C     183   267
C     193   289
;
proc glm data=info;
class group;
model MLW = group weight;
lsmeans group/alpha=0.05;
run;
quit;
```

### SAS Output (Part C: ANCOVA One-Way, Between-Subjects)

```
Example 35C: ANCOVA One-Way, Between-Subjects Design                                1

The GLM Procedure

   Class Level Information

Class         Levels    Values

group              3    A B C
```

```
Number of Observations Read          24
Number of Observations Used          24

Dependent Variable: MLW

                                   Sum of
Source                     DF         Squares     Mean Square    F Value    Pr > F

Model                       3     14023.05219      4674.35073       6.76    0.0025

Error                      20     13828.57281       691.42864

Corrected Total            23     27851.62500


R-Square     Coeff Var      Root MSE      MLW Mean

0.503491      9.279234      26.29503      283.3750


Source                     DF       Type I SS     Mean Square    F Value    Pr > F

group                       2      2889.25000      1444.62500       2.09    0.1500
weight                      1     11133.80219     11133.80219      16.10    0.0007


Source                     DF     Type III SS     Mean Square    F Value    Pr > F

group                       2      4912.02861      2456.01431       3.55    0.0479
weight                      1     11133.80219     11133.80219      16.10    0.0007

Least Squares Means

group      MLW LSMEAN

A          275.748163
B          303.668620
C          270.708217
```

### Output Explanation (Part C: ANCOVA One-Way, Between-Subjects)

The ANCOVA shows a significant difference among training groups on the maximum lifting weight since the p-value (0.0479) is less than the stated p-value (0.05) when adjusted for the covariate of soldier weight. The three training group means (i.e., A = 275.75, B = 303.67, and C = 270.71) are adjusted for the significant covariate, soldier weight. Post hoc analysis on the adjusted means for the three training conditions is needed to isolate these differences.

# Section 5. Empirical Model Building

## Example 36: Multiple Linear Regression

**(Click in this red rectangle to see the Reference Notes on Example 36.)**

### Problem

Location in Williges (2006) Table of Contents
Section 5, Topic 22. Multiple Regression, Part 22.2.3. Multiple Regression Example

Page(s) in Williges (2006) Reference Material: 732-736

Problem Description
The commander's combat operation performance in a battalion level command and control center for the Army is scored on a 100 point scale. Scores of fifteen battalion commanders are predicted as a function of four command and control tasks. The predictors are the time to complete Recognition, Decision, Communication, and, Evaluation tasks. What is the linear relationship of these four tasks on predicting the performance score? Are any of these predictors significant ($p < 0.05$)?

Context/Purpose
Determine the predictive relationship of Recognition, Decision, Communication, and Evaluation task completion times on a commander's combat operation performance score.

Statistical Decision Criteria
Conduct a multiple linear regression using the four task completion times as predictors of combat operation performance and test the significance of each partial regression weight at the 0.05 level of significance.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 36: Multiple Regression';
data info;
input rec dec com eval score;
lines;
56    47    59    55    76
60    49    57    53    80
59    50    64    57    86
52    55    52    54    75
51    45    55    58    66
54    58    53    60    76
60    49    57    62    90
57    50    54    53    71
58    53    56    54    77
53    57    53    56    79
63    45    54    51    83
54    53    55    50    70
58    50    58    55    76
60    50    57    55    75
55    56    50    53    73
;
```

```
proc glm data=info;
model score = rec dec com eval/XPX I;
run;
quit;
```

## SAS Output

```
Example 36: Multiple Regression                                                 1

The GLM Procedure

Number of Observations Read          15
Number of Observations Used          15

                                The X'X Matrix

                Intercept          rec          dec          com         eval        score

Intercept              15          850          767          834          826         1153
rec                   850        48334        43371        47332        46788        65532
dec                   767        43371        39453        42552        42247        58922
com                   834        47332        42552        46528        45959        64234
eval                  826        46788        42247        45959        45628        63592
score                1153        65532        58922        64234        63592        89159


                             X'X Inverse Matrix

                Intercept          rec          dec          com         eval        score

Intercept     97.010904214 -0.521298757 -0.560280652 -0.377620494 -0.322504682  -85.82673826
rec           -0.521298757  0.008647366  0.002120518 -0.003052919  0.001681497    1.3955088796
dec           -0.560280652  0.002120518  0.006343533  0.002973381 -0.000900147    0.4819552613
com           -0.377620494 -0.003052919  0.002973381  0.010132472 -0.002992466    0.28959255
eval          -0.322504682  0.001681497 -0.000900147 -0.002992466  0.007983567    0.7784971601
score        -85.82673826   1.395508879  0.481955261  0.28959255   0.7784971601 161.09415695
```

```
Dependent Variable: score

                                   Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F

Model                        4     370.6391764      92.6597941       5.75    0.0114

Error                       10     161.0941570      16.1094157

Corrected Total             14     531.7333333


R-Square     Coeff Var      Root MSE    score Mean

0.697040      5.221579      4.013654      76.86667


Source                      DF       Type I SS     Mean Square    F Value    Pr > F

rec                          1     228.0185923     228.0185923      14.15    0.0037
dec                          1      29.1946675      29.1946675       1.81    0.2080
com                          1      37.5127582      37.5127582       2.33    0.1580
eval                         1      75.9131583      75.9131583       4.71    0.0551


Source                      DF     Type III SS     Mean Square    F Value    Pr > F

rec                          1     225.2067152     225.2067152      13.98    0.0039
dec                          1      36.6169526      36.6169526       2.27    0.1626
com                          1       8.2767401       8.2767401       0.51    0.4899
eval                         1      75.9131583      75.9131583       4.71    0.0551


                                  Standard
Parameter         Estimate           Error    t Value    Pr > |t|

Intercept     -85.82673826     39.53212596      -2.17      0.0551
rec             1.39550888      0.37323454       3.74      0.0039
dec             0.48195526      0.31967268       1.51      0.1626
com             0.28959255      0.40401512       0.72      0.4899
eval            0.77849716      0.35862321       2.17      0.0551
```

### Output Explanation

The multiple linear regression equation of the four task completion times used to predict the overall combat performance score is: Performance Score = -85.83 + 1.40rec + 0.48dec + 0.29com + 0.78eval. The ANOVA on regression indicated that this multiple regression predicts a significant amount of combat performance score variance since the obtained p-value (0.0114) is less than the stated value (0.05). In addition, only the partial regression weight for Recognition task completion time is significant (p = 0.0039), given that the other three predictors are present in the multiple regression equation.

## Example 37: Best Regression Equation

**(Click in this red rectangle to see the Reference Notes on Example 37.)**

### Problem

Location in Williges (2006) Table of Contents
Section 5, Topic 22. Multiple Regression, Part 22.2.5. Best Equation Example

Page(s) in Williges (2006) Reference Material: 745-752

Problem Description
The commander's combat operation performance in a battalion level command and control center for the Army is scored on a 100 point scale. Scores of fifteen battalion commanders are predicted as a function of four command and control tasks. The predictors are the time to complete Recognition, Decision, Communication, and, Evaluation tasks. What is the best set of significant linear predictors to use in the prediction equation ($p < 0.05$)?

Context/Purpose
Determine the best subset of four task completion time predictors used to predict a commander's combat operation performance score.

Statistical Decision Criteria
Use a variety of classical and modern regression procedures to choose the overall best subset of four completion task times to use as predictors in the multiple linear regression equation.

### SAS Input (Part A. Backward Selection)

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 37A: Best Equation: Backward Selection';
data info;
input rec dec com eval score;
lines;
56    47    59    55    76
60    49    57    53    80
59    50    64    57    86
52    55    52    54    75
51    45    55    58    66
54    58    53    60    76
60    49    57    62    90
57    50    54    53    71
58    53    56    54    77
53    57    53    56    79
63    45    54    51    83
54    53    55    50    70
58    50    58    55    76
60    50    57    55    75
55    56    50    53    73
;
proc reg corr data=info;
model score = rec dec com eval/selection=b slstay=0.05 alpha=0.05;
```

```
run;
quit;
```

## SAS Output (Part A. Backward Selection)

```
Example 37A: Best Equation: Backward Selection                                    1

The REG Procedure

Number of Observations Read          15
Number of Observations Used          15


                                    Correlation

Variable              rec              dec              com             eval            score

rec                1.0000          -0.4669           0.4434          -0.1207           0.6548
dec               -0.4669           1.0000          -0.4856           0.0595          -0.0985
com                0.4434          -0.4856           1.0000           0.2225           0.4394
eval              -0.1207           0.0595           0.2225           1.0000           0.3632
score              0.6548          -0.0985           0.4394           0.3632           1.0000


Dependent Variable: score

Number of Observations Read          15
Number of Observations Used          15

Backward Elimination: Step 0


All Variables Entered: R-Square = 0.6970 and C(p) = 5.0000


                           Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F

Model                     4      370.63918       92.65979       5.75    0.0114
Error                    10      161.09416       16.10942
Corrected Total          14      531.73333


             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F

Intercept    -85.82674     39.53213     75.93197     4.71  0.0551
rec            1.39551      0.37323    225.20672    13.98  0.0039
dec            0.48196      0.31967     36.61695     2.27  0.1626
com            0.28959      0.40402      8.27674     0.51  0.4899
eval           0.77850      0.35862     75.91316     4.71  0.0551

Bounds on condition number: 1.5969, 22.671
------------------------------------------------------------------------------------------
```

```
Backward Elimination: Step 1

Variable com Removed: R-Square = 0.6815 and C(p) = 3.5138

                          Analysis of Variance

                                   Sum of           Mean
Source                   DF       Squares         Square    F Value    Pr > F

Model                     3     362.36244      120.78748       7.84    0.0045
Error                    11     169.37090       15.39735
Corrected Total          14     531.73333


Dependent Variable: score

Backward Elimination: Step 1

             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F

Intercept    -75.03410     35.73541     67.88376     4.41  0.0596
rec            1.48276      0.34494    284.51389    18.48  0.0013
dec            0.39697      0.29024     28.80436     1.87  0.1987
eval           0.86402      0.33063    105.14918     6.83  0.0241

Bounds on condition number: 1.2931, 10.76
------------------------------------------------------------------------------------------------

Backward Elimination: Step 2


Variable dec Removed: R-Square = 0.6273 and C(p) = 3.3018


                          Analysis of Variance

                                   Sum of           Mean
Source                   DF       Squares         Square    F Value    Pr > F

Model                     2     333.55808      166.77904      10.10    0.0027
Error                    12     198.17526       16.51460
Corrected Total          14     531.73333


             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F

Intercept    -42.42082     27.56566     39.11015     2.37  0.1498
rec            1.26389      0.31647    263.40870    15.95  0.0018
eval           0.86562      0.34242    105.53948     6.39  0.0265

Bounds on condition number: 1.0148, 4.0591
------------------------------------------------------------------------------------------------

All variables left in the model are significant at the 0.0500 level.
```

```
Dependent Variable: score

                        Summary of Backward Elimination

        Variable     Number     Partial      Model
Step    Removed      Vars In    R-Square     R-Square     C(p)      F Value    Pr > F

  1     com             3       0.0156       0.6815       3.5138       0.51    0.4899
  2     dec             2       0.0542       0.6273       3.3018       1.87    0.1987
```

**Output Explanation (Part A. Backward Selection)**

The backward selection begins with all of the parameters in the model and compares it to a specified level of 0.05. If the parameter is less than the specified value, it remains in the model and the next parameter is removed and the new model is tested. The best relationship between performance time and recognition, decision, communication, and evaluation tasks is explained by the following multiple linear regression model: Performance Score = -42.42 + 1.26rec + 0.86eval. The communication and decision task predictors were eliminated because the p-values (0.49 and 0.19) are greater than the criterion value (0.05). The p-value for the new model (0.0005) is statistically significant at the 0.05 level.

**SAS Input (Part B. Forward Selection)**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 37B: Best Equation: Forward Selection';
data info;
input rec dec com eval score;
lines;
56    47    59    55    76
60    49    57    53    80
59    50    64    57    86
52    55    52    54    75
51    45    55    58    66
54    58    53    60    76
60    49    57    62    90
57    50    54    53    71
58    53    56    54    77
53    57    53    56    79
63    45    54    51    83
54    53    55    50    70
58    50    58    55    76
60    50    57    55    75
55    56    50    53    73
;
proc reg corr data=info;
model score = rec dec com eval/selection=f slentry=0.05 alpha=0.05;
run;
quit;
```

## SAS Output (Part B. Forward Selection)

```
Example 37B: Best Equation: Forward Selection                                    1

The REG Procedure

Number of Observations Read          15
Number of Observations Used          15


                                    Correlation

Variable              rec              dec              com             eval            score

rec                1.0000          -0.4669           0.4434          -0.1207           0.6548
dec               -0.4669           1.0000          -0.4856           0.0595          -0.0985
com                0.4434          -0.4856           1.0000           0.2225           0.4394
eval              -0.1207           0.0595           0.2225           1.0000           0.3632
score              0.6548          -0.0985           0.4394           0.3632           1.0000

Forward Selection: Step 1

Variable rec Entered: R-Square = 0.4288 and C(p) = 7.8532

                           Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F

Model                     1      228.01859      228.01859       9.76    0.0081
Error                    13      303.71474       23.36267
Corrected Total          14      531.73333


            Parameter     Standard
Variable     Estimate        Error   Type II SS  F Value  Pr > F

Intercept    10.71793     21.21049      5.96544     0.26  0.6218
rec           1.16733      0.37365    228.01859     9.76  0.0081

Bounds on condition number: 1, 1
------------------------------------------------------------------------------------------

Forward Selection: Step 2

Variable eval Entered: R-Square = 0.6273 and C(p) = 3.3018

                           Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F

Model                     2      333.55808      166.77904      10.10    0.0027
Error                    12      198.17526       16.51460
Corrected Total          14      531.73333
```

```
Forward Selection: Step 2

             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F

Intercept    -42.42082     27.56566     39.11015     2.37  0.1498
rec            1.26389      0.31647    263.40870    15.95  0.0018
eval           0.86562      0.34242    105.53948     6.39  0.0265

Bounds on condition number: 1.0148, 4.0591
------------------------------------------------------------------------------------------------

No other variable met the 0.0500 significance level for entry into the model.


                          Summary of Forward Selection

        Variable     Number     Partial      Model
Step    Entered      Vars In    R-Square    R-Square     C(p)      F Value    Pr > F

  1     rec             1       0.4288      0.4288      7.8532       9.76    0.0081
  2     eval            2       0.1985      0.6273      3.3018       6.39    0.0265
```

**Output Explanation (Part B. Forward Selection)**

Forward selection begins with no parameters in the model. It adds one parameter and then compares the F-test to the specified value (0.05). If the F-test is less then the specified value, the parameter is added and the next one is tested. The parameters meeting this criterion are included in the final multiple regression model (Performance Score = -42.42 + 1.26rec + 0.87eval). The obtained p-value for this model (0.0027) is significant at the stated 0.05 level.

**SAS Input (Part C. Stepwise Selection)**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 37C: Best Equation: Stepwise Selection';
data info;
input rec dec com eval score;
lines;
56    47    59    55    76
60    49    57    53    80
59    50    64    57    86
52    55    52    54    75
51    45    55    58    66
54    58    53    60    76
60    49    57    62    90
57    50    54    53    71
58    53    56    54    77
53    57    53    56    79
63    45    54    51    83
```

```
54    53    55    50    70
58    50    58    55    76
60    50    57    55    75
55    56    50    53    73
;
proc reg corr data=info;
model score = rec dec com eval/selection=stepwise slstay=0.05 slentry=0.10
alpha=0.05;
run;
quit;
```

### SAS Output (Part C. Stepwise Selection)

```
Example 37C: Best Equation: Stepwise Selection                                                    1

The REG Procedure

Number of Observations Read          15
Number of Observations Used          15

                                      Correlation

Variable              rec              dec              com             eval            score

rec                1.0000          -0.4669           0.4434          -0.1207           0.6548
dec               -0.4669           1.0000          -0.4856           0.0595          -0.0985
com                0.4434          -0.4856           1.0000           0.2225           0.4394
eval              -0.1207           0.0595           0.2225           1.0000           0.3632
score              0.6548          -0.0985           0.4394           0.3632           1.0000


Stepwise Selection: Step 1


Variable rec Entered: R-Square = 0.4288 and C(p) = 7.8532


                            Analysis of Variance

                                     Sum of           Mean
Source                    DF        Squares         Square    F Value    Pr > F

Model                      1      228.01859      228.01859       9.76    0.0081
Error                     13      303.71474       23.36267
Corrected Total           14      531.73333


             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F

Intercept     10.71793     21.21049      5.96544     0.26  0.6218
rec            1.16733      0.37365    228.01859     9.76  0.0081

Bounds on condition number: 1, 1
--------------------------------------------------------------------------------------------
```

```
Stepwise Selection: Step 2

Variable eval Entered: R-Square = 0.6273 and C(p) = 3.3018

                            Analysis of Variance

                                     Sum of           Mean
Source                    DF        Squares         Square    F Value    Pr > F

Model                      2      333.55808      166.77904      10.10    0.0027
Error                     12      198.17526       16.51460
Corrected Total           14      531.73333


             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F

Intercept    -42.42082     27.56566     39.11015     2.37  0.1498
rec            1.26389      0.31647    263.40870    15.95  0.0018
eval           0.86562      0.34242    105.53948     6.39  0.0265

Bounds on condition number: 1.0148, 4.0591
------------------------------------------------------------------------------------------------


All variables left in the model are significant at the 0.0500 level.

No other variable met the 0.1000 significance level for entry into the model.

                            Summary of Stepwise Selection

      Variable    Variable    Number    Partial     Model
Step  Entered     Removed     Vars In   R-Square    R-Square    C(p)     F Value   Pr > F

  1   rec                        1       0.4288      0.4288     7.8532      9.76   0.0081
  2   eval                       2       0.1985      0.6273     3.3018      6.39   0.0265
```

### Output Explanation (Part C. Stepwise Selection)

The stepwise procedure first determines if the parameter should stay in the model and then determines other parameters that should be added. This iterative procedure continues until all the resulting subset of predictors are significant at the 0.05 level of significance. The best resulting multiple regression equation using the stepwise procedure is: Performance Score = -42.42 + 1.26rec + 0.87eval.

### SAS Input (Part D. All Possible Regressions)

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 37D: Best Equation: All Possible Regressions';
data info;
input rec dec com eval score;
lines;
56    47    59    55    76
```

```
60    49    57    53    80
59    50    64    57    86
52    55    52    54    75
51    45    55    58    66
54    58    53    60    76
60    49    57    62    90
57    50    54    53    71
58    53    56    54    77
53    57    53    56    79
63    45    54    51    83
54    53    55    50    70
58    50    58    55    76
60    50    57    55    75
55    56    50    53    73
;
proc reg corr data=info;
model score = rec dec com eval/selection=rsquare alpha=0.05;
run;
quit;
```

### SAS Output (Part D. All Possible Regressions)

```
Example 37D: Best Equation: All Possible Regressions                    1

The REG Procedure

Number of Observations Read          15
Number of Observations Used          15


                                  Correlation

Variable              rec              dec              com             eval            score

rec                1.0000          -0.4669           0.4434          -0.1207           0.6548
dec               -0.4669           1.0000          -0.4856           0.0595          -0.0985
com                0.4434          -0.4856           1.0000           0.2225           0.4394
eval              -0.1207           0.0595           0.2225           1.0000           0.3632
score              0.6548          -0.0985           0.4394           0.3632           1.0000
```

```
Number in
  Model      R-Square    Variables in Model

       1       0.4288    rec
       1       0.1931    com
       1       0.1319    eval
       1       0.0097    dec
-----------------------------------------
       2       0.6273    rec eval
       2       0.4837    rec dec
       2       0.4565    rec com
       2       0.2672    com eval
       2       0.2103    dec com
       2       0.1464    dec eval
-----------------------------------------
       3       0.6815    rec dec eval
       3       0.6282    rec com eval
       3       0.5543    rec dec com
       3       0.2735    dec com eval
-----------------------------------------
       4       0.6970    rec dec com eval
```

**Output Explanation (Part D. All Possible Regressions)**

The $R^2$ value for the model that includes all four parameters is the largest (0.70) thus indicating that all four should be included in the model. The relationship between combat performance as predicted by recognition, decision, communication, and evaluation task completion times is expressed in the following multiple linear regression equation: Performance Score = -85.83 + 1.40rec + 0.48dec + 0.30com + 0.78eval which is equivalent to the multiple regression conducted in Example 35.

**SAS Input (Part E. PRESS Statistic)**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 37E: Best Equation: PRESS Statistic';
data info;
input rec dec com eval score;
lines;
56    47    59    55    76
60    49    57    53    80
59    50    64    57    86
52    55    52    54    75
51    45    55    58    66
54    58    53    60    76
60    49    57    62    90
57    50    54    53    71
58    53    56    54    77
53    57    53    56    79
63    45    54    51    83
54    53    55    50    70
58    50    58    55    76
60    50    57    55    75
55    56    50    53    73
;
```

```
proc reg corr data=info;
model score = rec/p;
output Residual=Residual PRESS=PressRes;
model score = rec eval/p;
output Residual=Residual PRESS=PressRes;
model score = rec dec eval/p;
output Residual=Residual PRESS=PressRes;
model score = rec dec com eval/p;
output Residual=Residual PRESS=PressRes;
proc print;
run;
quit;
```

## SAS Output (Part E. PRESS Statistic)

```
Example 37E: Best Equation: PRESS Statistic                                                    1

The REG Procedure

Number of Observations Read          15
Number of Observations Used          15

                                     Correlation

Variable              rec            score             eval              dec              com

rec                1.0000           0.6548          -0.1207          -0.4669           0.4434
score              0.6548           1.0000           0.3632          -0.0985           0.4394
eval              -0.1207           0.3632           1.0000           0.0595           0.2225
dec               -0.4669          -0.0985           0.0595           1.0000          -0.4856
com                0.4434           0.4394           0.2225          -0.4856           1.0000

----------------------------------------------------------------------------------------------
Model: MODEL 1
                           Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F

Model                     1      228.01859      228.01859       9.76    0.0081
Error                    13      303.71474       23.36267
Corrected Total          14      531.73333


Root MSE              4.83349    R-Square     0.4288
Dependent Mean       76.86667    Adj R-Sq     0.3849
Coeff Var             6.28815


                      Parameter Estimates

                    Parameter       Standard
Variable     DF      Estimate          Error    t Value    Pr > |t|

Intercept     1      10.71793       21.21049       0.51      0.6218
rec           1       1.16733        0.37365       3.12      0.0081
```

Output Statistics

| Obs | Dependent Variable | Predicted Value | Residual |
|---|---|---|---|
| 1 | 76.0000 | 76.0884 | -0.0884 |
| 2 | 80.0000 | 80.7578 | -0.7578 |
| 3 | 86.0000 | 79.5904 | 6.4096 |
| 4 | 75.0000 | 71.4191 | 3.5809 |
| 5 | 66.0000 | 70.2518 | -4.2518 |
| 6 | 76.0000 | 73.7538 | 2.2462 |
| 7 | 90.0000 | 80.7578 | 9.2422 |
| 8 | 71.0000 | 77.2558 | -6.2558 |
| 9 | 77.0000 | 78.4231 | -1.4231 |
| 10 | 79.0000 | 72.5865 | 6.4135 |
| 11 | 83.0000 | 84.2598 | -1.2598 |
| 12 | 70.0000 | 73.7538 | -3.7538 |
| 13 | 76.0000 | 78.4231 | -2.4231 |
| 14 | 75.0000 | 80.7578 | -5.7578 |
| 15 | 73.0000 | 74.9211 | -1.9211 |

| | |
|---|---|
| Sum of Residuals | 0 |
| Sum of Squared Residuals | 303.71474 |
| Predicted Residual SS (PRESS) | 404.53009 |

---

Model: MODEL2

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | **2** | **333.55808** | **166.77904** | **10.10** | **0.0027** |
| Error | 12 | 198.17526 | 16.51460 | | |
| Corrected Total | 14 | 531.73333 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 4.06382 | **R-Square** | **0.6273** |
| Dependent Mean | 76.86667 | **Adj R-Sq** | **0.5652** |
| Coeff Var | 5.28684 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| **Intercept** | **1** | **-42.42082** | **27.56566** | **-1.54** | **0.1498** |
| **rec** | **1** | **1.26389** | **0.31647** | **3.99** | **0.0018** |
| **eval** | **1** | **0.86562** | **0.34242** | **2.53** | **0.0265** |

Model: MODEL2
Dependent Variable: score

Output Statistics

| Obs | Dependent Variable | Predicted Value | Residual |
|---|---|---|---|
| 1 | 76.0000 | 75.9664 | 0.0336 |
| 2 | 80.0000 | 79.2907 | 0.7093 |
| 3 | 86.0000 | 81.4893 | 4.5107 |
| 4 | 75.0000 | 70.0452 | 4.9548 |
| 5 | 66.0000 | 72.2438 | -6.2438 |
| 6 | 76.0000 | 77.7667 | -1.7667 |
| 7 | 90.0000 | 87.0813 | 2.9187 |
| 8 | 71.0000 | 75.4990 | -4.4990 |
| 9 | 77.0000 | 77.6285 | -0.6285 |
| 10 | 79.0000 | 73.0403 | 5.9597 |
| 11 | 83.0000 | 81.3511 | 1.6489 |
| 12 | 70.0000 | 69.1105 | 0.8895 |
| 13 | 76.0000 | 78.4942 | -2.4942 |
| 14 | 75.0000 | 81.0219 | -6.0219 |
| 15 | 73.0000 | 72.9712 | 0.0288 |

Sum of Residuals 0
Sum of Squared Residuals 198.17526
**Predicted Residual SS (PRESS) 326.63522**

---

Model: MODEL3
Dependent Variable: score

Number of Observations Read 15
Number of Observations Used 15

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 362.36244 | 120.78748 | 7.84 | 0.0045 |
| Error | 11 | 169.37090 | 15.39735 | | |
| Corrected Total | 14 | 531.73333 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 3.92395 | R-Square | 0.6815 |
| Dependent Mean | 76.86667 | Adj R-Sq | 0.5946 |
| Coeff Var | 5.10487 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -75.03410 | 35.73541 | -2.10 | 0.0596 |
| rec | 1 | 1.48276 | 0.34494 | 4.30 | 0.0013 |
| dec | 1 | 0.39697 | 0.29024 | 1.37 | 0.1987 |
| eval | 1 | 0.86402 | 0.33063 | 2.61 | 0.0241 |

Output Statistics

| Obs | Dependent Variable | Predicted Value | Residual |
|---|---|---|---|
| 1 | 76.0000 | 74.1797 | 1.8203 |
| 2 | 80.0000 | 79.1767 | 0.8233 |
| 3 | 86.0000 | 81.5470 | 4.4530 |
| 4 | 75.0000 | 70.5604 | 4.4396 |
| 5 | 66.0000 | 68.5640 | -2.5640 |
| 6 | 76.0000 | 79.9010 | -3.9010 |
| 7 | 90.0000 | 86.9529 | 3.0471 |
| 8 | 71.0000 | 75.1254 | -4.1254 |
| 9 | 77.0000 | 78.6631 | -1.6631 |
| 10 | 79.0000 | 74.5652 | 4.4348 |
| 11 | 83.0000 | 80.3090 | 2.6910 |
| 12 | 70.0000 | 69.2759 | 0.7241 |
| 13 | 76.0000 | 78.3362 | -2.3362 |
| 14 | 75.0000 | 81.3017 | -6.3017 |
| 15 | 73.0000 | 74.5417 | -1.5417 |

| | |
|---|---|
| Sum of Residuals | 0 |
| Sum of Squared Residuals | 169.37090 |
| Predicted Residual SS (PRESS) | 413.80891 |

---

Model: MODEL4

| | |
|---|---|
| Number of Observations Read | 15 |
| Number of Observations Used | 15 |

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 4 | 370.63918 | 92.65979 | 5.75 | 0.0114 |
| Error | 10 | 161.09416 | 16.10942 | | |
| Corrected Total | 14 | 531.73333 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 4.01365 | R-Square | 0.6970 |
| Dependent Mean | 76.86667 | Adj R-Sq | 0.5759 |
| Coeff Var | 5.22158 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -85.82674 | 39.53213 | -2.17 | 0.0551 |
| rec | 1 | 1.39551 | 0.37323 | 3.74 | 0.0039 |
| dec | 1 | 0.48196 | 0.31967 | 1.51 | 0.1626 |
| com | 1 | 0.28959 | 0.40402 | 0.72 | 0.4899 |
| eval | 1 | 0.77850 | 0.35862 | 2.17 | 0.0551 |

Output Statistics

| Obs | Dependent Variable | Predicted Value | Residual |
|---|---|---|---|
| 1 | 76.0000 | 74.8770 | 1.1230 |
| 2 | 80.0000 | 79.2867 | 0.7133 |
| 3 | 86.0000 | 83.5143 | 2.4857 |
| 4 | 75.0000 | 70.3449 | 4.6551 |
| 5 | 66.0000 | 68.1126 | -2.1126 |
| 6 | 76.0000 | 79.5424 | -3.5424 |
| 7 | 90.0000 | 86.2932 | 3.7068 |
| 8 | 71.0000 | 74.7134 | -3.7134 |
| 9 | 77.0000 | 78.9124 | -1.9124 |
| 10 | 79.0000 | 74.5509 | 4.4491 |
| 11 | 83.0000 | 79.1197 | 3.8803 |
| 12 | 70.0000 | 69.9268 | 0.0732 |
| 13 | 76.0000 | 78.8243 | -2.8243 |
| 14 | 75.0000 | 81.3257 | -6.3257 |
| 15 | 73.0000 | 73.6557 | -0.6557 |

| | |
|---|---|
| Sum of Residuals | 0 |
| Sum of Squared Residuals | 161.09416 |
| Predicted Residual SS (PRESS) | 504.76277 |

```
Example 36E: Best Equation: PRESS Statistic

Obs    rec    dec    com    eval    score    Residual    PressRes

  1     56     47     59     55       76      1.12304      1.4758
  2     60     49     57     53       80      0.71327      0.8445
  3     59     50     64     57       86      2.48569      6.2314
  4     52     55     52     54       75      4.65508      6.0419
  5     51     45     55     58       66     -2.11263    -10.1015
  6     54     58     53     60       76     -3.54238     -6.2746
  7     60     49     57     62       90      3.70680      8.5234
  8     57     50     54     53       71     -3.71338     -4.2290
  9     58     53     56     54       77     -1.91244     -2.1930
 10     53     57     53     56       79      4.44907      5.7759
 11     63     45     54     51       83      3.88034      9.5846
 12     54     53     55     50       70      0.07318      0.1154
 13     58     50     58     55       76     -2.82425     -3.1626
 14     60     50     57     55       75     -6.32568     -7.3258
 15     55     56     50     53       73     -0.65572     -0.9364
```

**Output Explanation (Part E. PRESS Statistic)**

The PRESS statistic is used to evaluate each possible variable and determine which should be included in the model by comparing how well it will predict the observed scores. The model with the smallest PRESS statistic should be selected for use. In this example the second model is the best equation since the PRESS statistic (326.64) is the smallest. The preferred model is Performance Score = -42.42 + 1.26rec + 0.87eval with a p-value (0.0027) less than 0.05, indicating that it is statistically significant.

**SAS Input (Part F. Mallows Cp)**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 37F: Best Equation: Mallows C(p)';
data info;
input rec dec com eval score;
lines;
56    47    59    55    76
60    49    57    53    80
59    50    64    57    86
52    55    52    54    75
51    45    55    58    66
54    58    53    60    76
60    49    57    62    90
57    50    54    53    71
58    53    56    54    77
53    57    53    56    79
63    45    54    51    83
54    53    55    50    70
58    50    58    55    76
60    50    57    55    75
55    56    50    53    73
;
```

```
proc reg corr data=info;
model score = rec dec com eval/selection=cp alpha=0.05;
run;
quit;
```

### SAS Output (Part F. Mallows Cp)

```
Example 37F: Best Equation: Mallows C(p)                                    1

The REG Procedure

Number of Observations Read          15
Number of Observations Used          15


                                    Correlation

Variable              rec              dec              com             eval            score

rec                1.0000          -0.4669           0.4434          -0.1207           0.6548
dec               -0.4669           1.0000          -0.4856           0.0595          -0.0985
com                0.4434          -0.4856           1.0000           0.2225           0.4394
eval              -0.1207           0.0595           0.2225           1.0000           0.3632
score              0.6548          -0.0985           0.4394           0.3632           1.0000


C(p) Selection Method

Number of Observations Read          15
Number of Observations Used          15



Number in
  Model          C(p)    R-Square    Variables in Model

      2        3.3018      0.6273    rec eval
      3        3.5138      0.6815    rec dec eval
      4        5.0000      0.6970    rec dec com eval
      3        5.2730      0.6282    rec com eval
      3        7.7123      0.5543    rec dec com
      1        7.8532      0.4288    rec
      2        8.0410      0.4837    rec dec
      2        8.9403      0.4565    rec com
      2       15.1880      0.2672    com eval
      1       15.6347      0.1931    com
      3       16.9798      0.2735    dec com eval
      2       17.0650      0.2103    dec com
      1       17.6531      0.1319    eval
      2       19.1751      0.1464    dec eval
      1       21.6872      0.0097    dec
```

**Output Explanation (Part F. Mallows Cp)**
Each of the parameters is added to the model and the Mallows Cp value is calculated. The model with the Mallows Cp value that is closest to the number of parameters minus one (p-1) is selected as best fit. For this example, three parameters (4-1=3) are closest to the Mallows Cp value of 3.30 from the second model. Consequently, the resulting best multiple linear regression model is: Performance Score = -42.42 + 1.26rec + 0.87eval.

**Overall Choice of Best Regression Equation**

The multiple linear regression equation of Performance Score = -42.42 + 1.26rec + 0.87eval was selected as the best equation using the forward selection, backward selection, stepwise, PRESS statistic, and the Mallow's Cp tests. The all possible selection method included all four variables in the best equation (Performance Score = -85.83 + 1.40rec + 0.48dec + 0.30com + 0.78eval). Based on all these procedures, the overall best consensus multiple linear regression equation seems to be the regression with only two predictors: Performance Score = -42.42 + 1.26rec + 0.87eval.

## Example 38: Polynomial Regression

### (Click in this red rectangle to see the Reference Notes on Example 38.)

**Problem**

Location in Williges (2006) Table of Contents
Section 5, Topic 22. Multiple Regression, Part 22.3.2. Polynomial Regression Example

Page(s) in Williges (2006) Reference Material: 756-763

Problem Description
A between-subjects experiment (n = 4) was conducted to build an empirical model of soldier percent reading comprehension of text presented on computer displays as a function of possible first- and second-order effects involving two different sizes of computer monitors (17 and 21 inch) and three different font sizes (12, 16, and 18 point). What is the resulting second-order model and were any first- and second-order parameters significant predictors ($p < 0.01$)?

Context/Purpose
Determine an empirical model that includes first and second order parameters to predict reading comprehension as a function of monitor size and font size.

Statistical Decision Criteria
Use polynomial regression to generate the empirical model and test the significance of the partial regression weights included in the model at the 0.01 level of significance.

**SAS Input (Part A. Two-Factor ANOVA)**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 38A: Polynomial Regression: Two-Factor ANOVA';
data six;
input Subject Monitor Font Reading;
lines;
1 17 12 35
2 17 12 42
3 17 12 39
4 17 12 40
5 21 12 50
6 21 12 47
7 21 12 49
8 21 12 52
9 17 16 39
10 17 16 44
11 17 16 38
12 17 16 45
13 21 16 49
14 21 16 52
15 21 16 54
16 21 16 48
17 17 18 47
18 17 18 46
```

```
19 17 18 50
20 17 18 44
21 21 18 46
22 21 18 50
23 21 18 49
24 21 18 47
;
proc glm data=six;
class Monitor Font Subject;
model Reading = Monitor Font Monitor*Font;
means Monitor Font Monitor*Font;
run;
quit;
```

### SAS Output (Part A. Two-Factor ANOVA)

```
Example 38A: Polynomial Regression: Two-Factor ANOVA                                  1

The GLM Procedure

                              Class Level Information

Class         Levels    Values

Monitor            2    17 21

Font               3    12 16 18

Subject           24    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24


Number of Observations Read          24
Number of Observations Used          24

                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F

Model                        5     434.3333333      86.8666667      12.26    <.0001

Error                       18     127.5000000       7.0833333

Corrected Total             23     561.8333333


R-Square     Coeff Var      Root MSE    Reading Mean

0.773064      5.796268      2.661453        45.91667


Source                      DF       Type I SS     Mean Square    F Value    Pr > F

Monitor                      1     294.0000000     294.0000000      41.51    <.0001
Font                         2      39.5833333      19.7916667       2.79    0.0877
Monitor*Font                 2     100.7500000      50.3750000       7.11    0.0053


Source                      DF     Type III SS     Mean Square    F Value    Pr > F
```

| | | | | | |
|---|---|---|---|---|---|
| **Monitor** | **1** | **294.0000000** | **294.0000000** | **41.51** | **<.0001** |
| Font | 2 | 39.5833333 | 19.7916667 | 2.79 | 0.0877 |
| **Monitor*Font** | **2** | **100.7500000** | **50.3750000** | **7.11** | **0.0053** |

| Level of Monitor | N | Reading Mean | Reading Std Dev |
|---|---|---|---|
| 17 | 12 | 42.4166667 | 4.33711956 |
| 21 | 12 | 49.4166667 | 2.35326981 |

| Level of Font | N | Reading Mean | Reading Std Dev |
|---|---|---|---|
| 12 | 8 | 44.2500000 | 6.08863109 |
| 16 | 8 | 46.1250000 | 5.74300817 |
| 18 | 8 | 47.3750000 | 2.13390989 |

| Level of Monitor | Level of Font | N | Reading Mean | Reading Std Dev |
|---|---|---|---|---|
| 17 | 12 | 4 | 39.0000000 | 2.94392029 |
| 17 | 16 | 4 | 41.5000000 | 3.51188458 |
| 17 | 18 | 4 | 46.7500000 | 2.50000000 |
| 21 | 12 | 4 | 49.5000000 | 2.08166600 |
| 21 | 16 | 4 | 50.7500000 | 2.75378527 |
| 21 | 18 | 4 | 48.0000000 | 1.82574186 |

**Output Explanation (Part A. Two-Factor ANOVA)**
A two-factor ANOVA was conducted to determine if there is a significant difference between monitor size and font size, as well as if there is any significance due to the interaction of these parameters. The analysis indicates that the monitor size has a significant effect on reading comprehension since the p-value (<0.0001) is less than the stated value (0.01). There is also an effect due to the interaction of monitor and font size (p-value = 0.0053). Post hoc analysis is required to determine which levels of monitor and font size interaction have an effect on reading comprehension. Subsequently, the data from this factorial design is used to generate an empirical model of these effects using polynomial regression.

**SAS Input (Part B. Complete Model)**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 38B: Polynomial Regression: Complete Model';
data six;
input Subject Monitor Font Reading;
lines;
1 17 12 35
2 17 12 42
3 17 12 39
4 17 12 40
```

```
5 21 12 50
6 21 12 47
7 21 12 49
8 21 12 52
9 17 16 39
10 17 16 44
11 17 16 38
12 17 16 45
13 21 16 49
14 21 16 52
15 21 16 54
16 21 16 48
17 17 18 47
18 17 18 46
19 17 18 50
20 17 18 44
21 21 18 46
22 21 18 50
23 21 18 49
24 21 18 47
;
proc glm data=six;
model Reading = Monitor Font Font*Font Monitor*Font Monitor*Font*Font;
run;
quit;
```

### SAS Output (Part B. Complete Model)

```
Example 38B: Polynomial Regression: Complete Model                                1

The GLM Procedure

Number of Observations Read          24
Number of Observations Used          24

                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F

Model                        5     434.3333333      86.8666667      12.26    <.0001

Error                       18     127.5000000       7.0833333

Corrected Total             23     561.8333333


R-Square     Coeff Var      Root MSE    Reading Mean

0.773064      5.796268      2.661453        45.91667


Source                      DF       Type I SS     Mean Square    F Value    Pr > F

Monitor                      1     294.0000000     294.0000000      41.51    <.0001
Font                         1      39.3601190      39.3601190       5.56    0.0299
Font*Font                    1       0.2232143       0.2232143       0.03    0.8611
Monitor*Font                 1      69.6696429      69.6696429       9.84    0.0057
Monitor*Font*Font            1      31.0803571      31.0803571       4.39    0.0506
```

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Monitor | 1 | 19.06831267 | 19.06831267 | 2.69 | 0.1182 |
| Font | 1 | 26.41791445 | 26.41791445 | 3.73 | 0.0694 |
| Font*Font | 1 | 31.29063112 | 31.29063112 | 4.42 | 0.0499 |
| Monitor*Font | 1 | 26.53812944 | 26.53812944 | 3.75 | 0.0688 |
| Monitor*Font*Font | 1 | 31.08035714 | 31.08035714 | 4.39 | 0.0506 |

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| **Intercept** | **536.4375000** | 302.0210784 | 1.78 | 0.0926 |
| **Monitor** | **-25.9375000** | 15.8085058 | -1.64 | 0.1182 |
| **Font** | **-80.5156250** | 41.6917134 | -1.93 | 0.0694 |
| **Font*Font** | **2.9453125** | 1.4013385 | 2.10 | 0.0499 |
| **Monitor*Font** | **4.2239583** | 2.1822440 | 1.94 | 0.0688 |
| **Monitor*Font*Font** | **-0.1536458** | 0.0733494 | -2.09 | 0.0506 |

### Output Explanation (Part B. Complete Model)

The prediction of reading comprehension as a function of monitor size (M) and font size (F) is explained by the complete polynomial regression model: Reading Comprehension = 536.43 – 25.94M – 80.51F + 2.95$F^2$ + 4.22MF – 0.15$MF^2$. This model is statistically significant since the p-value (>0.001) is less than the stated level (0.01) as determined by the ANOVA on regression. Again, there is a significant effect due to the size of the monitor (p-value <0.001) and the linear by linear component of the interaction of monitor and font size (p-value = 0.0057). The $MF^2$partial regression weight is a third-order effect that should be removed if only a model involving first and second-order effects is needed.

### SAS Input (Part C. Second-Order Model)

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 38C: Polynomial Regression: Second-Order Model';
data six;
input Subject Monitor Font Reading;
lines;
1 17 12 35
2 17 12 42
3 17 12 39
4 17 12 40
5 21 12 50
6 21 12 47
7 21 12 49
8 21 12 52
9 17 16 39
10 17 16 44
11 17 16 38
12 17 16 45
13 21 16 49
14 21 16 52
15 21 16 54
```

```
16 21 16 48
17 17 18 47
18 17 18 46
19 17 18 50
20 17 18 44
21 21 18 46
22 21 18 50
23 21 18 49
24 21 18 47
;
proc glm data=six;
model Reading = Monitor Font Font*Font Monitor*Font;
run;
quit;
```

### SAS Output (Part C. Second-Order Model)

```
Example 38C: Polynomial Regression: Second-Order Model                                1

The GLM Procedure

Number of Observations Read          24
Number of Observations Used          24

                                   Sum of
Source                     DF         Squares     Mean Square    F Value    Pr > F

Model                       4     403.2529762     100.8132440      12.08    <.0001

Error                      19     158.5803571       8.3463346

Corrected Total            23     561.8333333


R-Square     Coeff Var      Root MSE    Reading Mean

0.717745      6.291838      2.889002        45.91667


Source                     DF       Type I SS     Mean Square    F Value    Pr > F

Monitor                     1     294.0000000     294.0000000      35.23    <.0001
Font                        1      39.3601190      39.3601190       4.72    0.0428
Font*Font                   1       0.2232143       0.2232143       0.03    0.8718
Monitor*Font                1      69.6696429      69.6696429       8.35    0.0094


Source                     DF     Type III SS     Mean Square    F Value    Pr > F

Monitor                     1     120.8181235     120.8181235      14.48    0.0012
Font                        1      11.7784128      11.7784128       1.41    0.2495
Font*Font                   1       0.2232143       0.2232143       0.03    0.8718
Monitor*Font                1      69.6696429      69.6696429       8.35    0.0094
```

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| **Intercept** | **-89.12053571** | 48.94070999 | -1.82 | 0.0844 |
| **Monitor** | **6.98660714** | 1.83631920 | 3.80 | 0.0012 |
| **Font** | **6.22842262** | 5.24303290 | 1.19 | 0.2495 |
| **Font*Font** | **0.02604167** | 0.15924129 | 0.16 | 0.8718 |
| **Monitor*Font** | **-0.34151786** | 0.11820600 | -2.89 | 0.0094 |

**Output Explanation (Part C. Second-Order Model)**

The resulting polynomial regression model predicting reading comprehension as a function of first- and second-order effects of monitor size (M), font size (F) is: Reading Comprehension = -89.12 + 6.99M +6.23F + 0.03$F^2$ – 0.34MF. This model is statistically significant since the p < 0.0001 level is less than the stated level (p < 0.01). Again, there is a significant effect due to the size of the monitor (p = 0.0012) and the interaction of monitor and font size (p = 0.0094). Note that the $R^2$ coefficient of determination decreases as the order of the model decreases from the previous complete polynomial regression model and the beta weight change due to covariance among higher-order effects.

**SAS Input (Part D. Lack of Fit)**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 38D: Polynomial Regression: Lack of Fit';
data six;
input Subject Monitor Font Reading;
lines;
1 17 12 35
2 17 12 42
3 17 12 39
4 17 12 40
5 21 12 50
6 21 12 47
7 21 12 49
8 21 12 52
9 17 16 39
10 17 16 44
11 17 16 38
12 17 16 45
13 21 16 49
14 21 16 52
15 21 16 54
16 21 16 48
17 17 18 47
18 17 18 46
19 17 18 50
20 17 18 44
21 21 18 46
22 21 18 50
23 21 18 49
24 21 18 47
```

```
;
proc rsreg data=six;
model Reading = Monitor Font /lackfit;
run;
quit;
```

### SAS Output (Part D. Lack of Fit)

```
Example 38D: Polynomial Regression: Lack of Fit                                1

The RSREG Procedure

Coding Coefficients for the Independent Variables

Factor     Subtracted off      Divided by

Monitor         19.000000        2.000000
Font            15.000000        3.000000


 Response Surface for Variable Reading

Response Mean                  45.916667
Root MSE                        2.889002
R-Square                          0.7177
Coefficient of Variation          6.2918


                          Type I Sum
Regression          DF    of Squares    R-Square    F Value    Pr > F

Linear               2    333.360119      0.5933      19.97    <.0001
Quadratic            1      0.223214      0.0004       0.03    0.8718
Crossproduct         1     69.669643      0.1240       8.35    0.0094
Total Model          4    403.252976      0.7177      12.08    <.0001


                             Sum of
Residual            DF      Squares     Mean Square    F Value    Pr > F

Lack of Fit          1     31.080357     31.080357       4.39    0.0506
Pure Error          18    127.500000      7.083333
Total Error         19    158.580357      8.346335


                                                                  Parameter
                                                                   Estimate
                                    Standard                      from Coded
Parameter           DF    Estimate     Error    t Value    Pr > |t|     Data

Intercept            1   -89.120536  48.940710      -1.82      0.0844  45.578125
Monitor              1     6.986607   1.836319       3.80      0.0012   3.727679
Font                 1     6.228423   5.243033       1.19      0.2495   1.562500
Monitor*Monitor      0            0          .          .           .          0
Font*Monitor         1    -0.341518   0.118206      -2.89      0.0094  -2.049107
Font*Font            1     0.026042   0.159241       0.16      0.8718   0.234375
```

| Factor | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Monitor | 2 | 363.669643 | 181.834821 | 21.79 | <.0001 |
| Font | 3 | 109.252976 | 36.417659 | 4.36 | 0.0169 |

Canonical Analysis of Response Surface Based on Coded Data

| Factor | Critical Value Coded | Uncoded |
|---|---|---|
| Monitor | 1.178678 | 21.357355 |
| Font | 1.819172 | 20.457516 |

Predicted value at stationary point: 49.196219

| Eigenvalues | Eigenvectors Monitor | Font |
|---|---|---|
| 1.148421 | -0.665718 | 0.746203 |
| -0.914046 | 0.746203 | 0.665718 |

Stationary point is a saddle point.

**Output Explanation (Part D. Lack of Fit)**

The second-order model is exactly the same as calculated in Part C. An ANOVA on regression shows that the composite of the two linear components of the model ($p < 0.001$) is significant at the 0.01 level, and that the cross product model is also significant ($p = 0.0094$). The error due to lack of fit ($p = 0.051$) which is a third-order component is not significant ($p>0.01$) in the polynomial model.

**Summary of Polynomial Regression Example**

Shown below a partially revised ANOVA summary table that uses the information provided by the SAS output. The complete revised ANOVA summary table for this design can be found in the Williges (2006) reference. The empirical model describing reading comprehension is: Reading Comprehension = $-89.12 + 6.99M + 6.23F - 0.03F^2 - 0.34MF$ as shown in Parts C and D.

**Revised ANOVA Summary Table (Second-Order Empirical Model)**

| Source | DF | Type I Sum of Squares | F Value | Pr > F |
|---|---|---|---|---|
| Model | (4) | (403.252976) | (12.08) | (<.0001) |
| Monitor | 1 | 294.000000 | 35.23 | <.0001 |
| Font | 1 | 39.360119 | 4.72 | 0.0428 |
| Font*Font | 1 | 0.223214 | 0.03 | 0.8718 |
| Monitor*Font | 1 | 69.669643 | 8.35 | 0.0094 |
| Error | (19) | (158.58036) | | |
| Lack of Fit | 1 | 31.080357 | 4.39 | 0.0506 |
| Pure Error | 18 | 127.500000 | | |
| Corrected Total | 23 | 561.8333333 | | |

## Example 39: Orthogonal, Between-Subjects, Central-Composite Design

**(Click in this red rectangle to see the Reference Notes on Example 39.)**

### Problem

Location in Williges (2006) Table of Contents
Section 5, Topic 23. Central-Composite Designs (CCD), Part 23.4.1. Between-Subjects Example

Page(s) in Williges (2006) Reference Material: 796-801

Problem Description
A computer-generated Army surveillance display is tested to predict the effects of three target characteristics on the probability of target detection. The three parameters of interest are target size, target density, and target velocity. Forty-five soldiers were tested in a between-subjects, orthogonal, central-composite design. Is the complete orthogonal, second-order empirical model significant ($p < 0.05$)? Which predictors are significant and do significant higher-order predictors exist ($p < 0.05$)?

Context/Purpose
Develop a complete second-order empirical model that predicts the probability of target detection as a function of target size, target density, and target velocity.

Statistical Decision Criteria
Use an orthogonal, second-order, between-subjects central-composite design to develop the polynomial regression model and conduct an ANOVA on regression to test for significance at the 0.05 level. The coded value of $\alpha$ is set at ±1.216 to keep the partial regression weight orthogonal in the second-order empirical model.

### SAS Input (Part A. Between-Subjects Coded)

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 39A: Orthogonal, Between-Subjects-Subjects, Central-Composite
Design (Coded)';
data info;
input Subject Size Density Velocity Probability;
lines;
1      1          -1           1          0.70
2      1          -1           1          0.82
3      1          -1           1          0.78
4      1           1          -1          0.63
5      1           1          -1          0.44
6      1           1          -1          0.52
7     -1           1           1          0.65
8     -1           1           1          0.67
9     -1           1           1          0.86
10    -1          -1          -1          0.30
11    -1          -1          -1          0.45
```

```
12    -1          -1          -1          0.26
13    -1           1          -1          0.49
14    -1           1          -1          0.58
15    -1           1          -1          0.47
16    -1          -1           1          0.48
17    -1          -1           1          0.56
18    -1          -1           1          0.35
19     1          -1          -1          0.53
20     1          -1          -1          0.74
21     1          -1          -1          0.63
22     1           1           1          0.85
23     1           1           1          0.98
24     1           1           1          0.81
25    -1.216       0           0          0.36
26    -1.216       0           0          0.47
27    -1.216       0           0          0.55
28     0          -1.216       0          0.53
29     0          -1.216       0          0.74
30     0          -1.216       0          0.60
31     0           0          -1.216      0.58
32     0           0          -1.216      0.35
33     0           0          -1.216      0.25
34     1.216       0           0          0.77
35     1.216       0           0          0.93
36     1.216       0           0          0.81
37     0           1.216       0          0.62
38     0           1.216       0          0.93
39     0           1.216       0          0.68
40     0           0           1.216      0.86
41     0           0           1.216      0.94
42     0           0           1.216      0.96
43     0           0           0          0.75
43     0           0           0          0.73
45     0           0           0          0.62
;
proc glm data=info;
model Probability= Size Density Velocity Size*Density Size*Velocity
Density*Velocity Size*Size Density*Density Velocity*Velocity;
proc rsreg data=info;
model Probability= Size Density Velocity/LACKFIT;
run;
quit;
```

### SAS Output (Part A. Between-Subjects Coded)

```
Example 39A: Orthogonal, Between-Subjects-Subjects, Central-Composite Design (Coded)           1

The GLM Procedure

Number of Observations Read          45
Number of Observations Used          45
Dependent Variable: Probability
```

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | **9** | **1.25816155** | **0.13979573** | **11.26** | **<.0001** |
| Error | 35 | 0.43456290 | 0.01241608 | | |
| Corrected Total | 44 | 1.69272444 | | | |

| R-Square | Coeff Var | Root MSE | Probability Mean |
|---|---|---|---|
| 0.743276 | 17.54456 | 0.111427 | 0.635111 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Size | 1 | 0.41288853 | 0.41288853 | 33.25 | <.0001 |
| Density | 1 | 0.09722840 | 0.09722840 | 7.83 | 0.0083 |
| Velocity | 1 | 0.58662015 | 0.58662015 | 47.25 | <.0001 |
| Size*Density | 1 | 0.06933750 | 0.06933750 | 5.58 | 0.0238 |
| Size*Velocity | 1 | 0.00770417 | 0.00770417 | 0.62 | 0.4362 |
| Density*Velocity | 1 | 0.03450417 | 0.03450417 | 2.78 | 0.1044 |
| Size*Size | 1 | 0.02525858 | 0.02525858 | 2.03 | 0.1626 |
| Density*Density | 1 | 0.00537412 | 0.00537412 | 0.43 | 0.5149 |
| Velocity*Velocity | 1 | 0.01924592 | 0.01924592 | 1.55 | 0.2214 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Size** | **1** | **0.41288853** | **0.41288853** | **33.25** | **<.0001** |
| **Density** | **1** | **0.09722840** | **0.09722840** | **7.83** | **0.0083** |
| **Velocity** | **1** | **0.58662015** | **0.58662015** | **47.25** | **<.0001** |
| **Size*Density** | **1** | **0.06933750** | **0.06933750** | **5.58** | **0.0238** |
| Size*Velocity | 1 | 0.00770417 | 0.00770417 | 0.62 | 0.4362 |
| Density*Velocity | 1 | 0.03450417 | 0.03450417 | 2.78 | 0.1044 |
| Size*Size | 1 | 0.02532309 | 0.02532309 | 2.04 | 0.1621 |
| Density*Density | 1 | 0.00539361 | 0.00539361 | 0.43 | 0.5141 |
| Velocity*Velocity | 1 | 0.01924592 | 0.01924592 | 1.55 | 0.2214 |

Dependent Variable: Probability

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 0.7100326514 | 0.04237199 | 16.76 | <.0001 |
| Size | 0.1120737154 | 0.01943478 | 5.77 | <.0001 |
| Density | 0.0543856011 | 0.01943478 | 2.80 | 0.0083 |
| Velocity | 0.1335875076 | 0.01943478 | 6.87 | <.0001 |
| Size*Density | -.0537500000 | 0.02274504 | -2.36 | 0.0238 |
| Size*Velocity | 0.0179166667 | 0.02274504 | 0.79 | 0.4362 |
| Density*Velocity | 0.0379166667 | 0.02274504 | 1.67 | 0.1044 |
| Size*Size | -.0439565503 | 0.03077922 | -1.43 | 0.1621 |
| Density*Density | -.0202864066 | 0.03077922 | -0.66 | 0.5141 |
| Velocity*Velocity | -.0383208018 | 0.03077922 | -1.25 | 0.2214 |

The RSREG Procedure

Coding Coefficients for the Independent Variables

| Factor | Subtracted off | Divided by |
|---|---|---|
| Size | 0 | 1.216000 |
| Density | 0 | 1.216000 |
| Velocity | 0 | 1.216000 |

Response Surface for Variable Probability

| | |
|---|---|
| Response Mean | 0.635111 |
| Root MSE | 0.111427 |
| **R-Square** | **0.7433** |
| Coefficient of Variation | 17.5446 |

| Regression | DF | Type I Sum of Squares | R-Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Linear | 3 | 1.096737 | 0.6479 | 29.44 | <.0001 |
| Quadratic | 3 | 0.049879 | 0.0295 | 1.34 | 0.2774 |
| Crossproduct | 3 | 0.111546 | 0.0659 | 2.99 | 0.0438 |
| Total Model | 9 | 1.258162 | 0.7433 | 11.26 | <.0001 |

| Residual | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Lack of Fit** | **5** | **0.113096** | **0.022619** | **2.11** | **0.0915** |
| **Pure Error** | **30** | **0.321467** | **0.010716** | | |
| **Total Error** | **35** | **0.434563** | **0.012416** | | |

| Parameter | DF | Estimate | Standard Error | t Value | Pr > \|t\| | Parameter Estimate from Coded Data |
|---|---|---|---|---|---|---|
| **Intercept** | **1** | **0.710033** | 0.042372 | 16.76 | <.0001 | 0.710033 |
| **Size** | **1** | **0.112074** | 0.019435 | 5.77 | <.0001 | 0.136282 |
| **Density** | **1** | **0.054386** | 0.019435 | 2.80 | 0.0083 | 0.066133 |
| **Velocity** | **1** | **0.133588** | 0.019435 | 6.87 | <.0001 | 0.162442 |
| **Size*Size** | **1** | **-0.043957** | 0.030779 | -1.43 | 0.1621 | -0.064997 |
| **Density*Size** | **1** | **-0.053750** | 0.022745 | -2.36 | 0.0238 | -0.079478 |
| **Density*Density** | **1** | **-0.020286** | 0.030779 | -0.66 | 0.5141 | -0.029997 |
| **Velocity*Size** | **1** | **0.017917** | 0.022745 | 0.79 | 0.4362 | 0.026493 |
| **Velocity*Density** | **1** | **0.037917** | 0.022745 | 1.67 | 0.1044 | 0.056066 |
| **Velocity*Velocity** | **1** | **-0.038321** | 0.030779 | -1.25 | 0.2214 | -0.056663 |

| Factor | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Size | 4 | 0.515253 | 0.128813 | 10.37 | <.0001 |
| Density | 4 | 0.206464 | 0.051616 | 4.16 | 0.0074 |
| Velocity | 4 | 0.648074 | 0.162019 | 13.05 | <.0001 |

```
 Canonical Analysis of Response Surface Based on Coded Data

                  Critical Value
Factor             Coded         Uncoded

Size             9.961250       12.112880
Density        -15.954920      -19.401183
Velocity        -4.131266       -5.023620

Predicted value at stationary point: 0.525681


                                 Eigenvectors
 Eigenvalues            Size         Density        Velocity

    0.001174        -0.444288        0.841895        0.306302
   -0.046952         0.578941        0.008886        0.815321
   -0.105878         0.683693        0.539568       -0.491355

             Stationary point is a saddle point.
```

**Output Explanation (Part A. Between-Subjects Coded)**

By using coded-values of levels in the polynomial regression, the complete, second-order, empirical model that predicts the probability of target detection (P) as a function of the three display variables is: P = 0.7100 + 0.1121(Size) + 0.0544(Density) + 0.1336(Velocity) – 0.0538(Size x Density) + 0.0179(Size x Velocity) + 0.0379(Density x Velocity) - 0.0440($Size^2$) – 0.0203($Density^2$) - 0.0383($Velocity^2$). The p-value for the model (<0.001) is less than the specified significance level (0.05). Therefore, the relationship describing the decision rate is statistically significant. The $R^2$ value (0.74) indicates that approximately 74% of the variation in probability of target detection is accounted for by the second-order empirical model. The predictors for target size, density, and velocity are all significant at the 0.05 level and the linear by linear predictor of size x density is also significant in this model (p = 0.024). Shown below is a partially revised ANOVA summary table that uses the information provided by SAS. The complete ANOVA summary table for this design can be found in the Williges (2006).

**CCD Revised ANOVA Summary Table (Orthogonal Coded Between-Subjects Design)**

| Source | DF | Type III Sum of Squares | F Value | Pr > F |
|---|---|---|---|---|
| Model | (9) | (1.25816155) | (11.26) | (<.0001) |
| Size | 1 | 0.41288853 | 33.25 | <.0001 |
| Density | 1 | 0.09722840 | 7.83 | 0.0083 |
| Velocity | 1 | 0.58662015 | 47.25 | <.0001 |
| Size*Density | 1 | 0.06933750 | 5.58 | 0.0238 |
| Size*Velocity | 1 | 0.00770417 | 0.62 | 0.4362 |
| Density*Velocity | 1 | 0.03450417 | 2.78 | 0.1044 |
| Size*Size | 1 | 0.02532309 | 2.04 | 0.1621 |
| Density*Density | 1 | 0.00539361 | 0.43 | 0.5141 |
| Velocity*Velocity | 1 | 0.01924592 | 1.55 | 0.2214 |
| Total Error | (35) | (0.434563) | | |
| Lack of Fit | 5 | 0.113096 | 2.11 | 0.0915 |
| Pure Error | 30 | 0.321467 | | |
| Corrected Total | 44 | 1.692724 | | |

**SAS Input (Part B. Between-Subjects Raw Scores)**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

| Level | Target Size (pixels) | Target Density (# per hour) | Target Velocity (Km per hour) |
|---|---|---|---|
| **-1.216** | 11 | 11 | 8 |
| **-1** | 12 | 12 | 10 |
| **0** | 18 | 16 | 20 |
| **1** | 24 | 20 | 30 |
| **1.216** | 25 | 21 | 32 |

```
options nodate nocenter pageno=1;
title'Example 39B: Orthogonal, Between-Subjects-Subjects, Central-Composite
Design (Raw Score)';
data info;
input Subject Size Density Velocity Probability;
lines;
1     24    12    30          0.70
2     24    12    30          0.82
3     24    12    30          0.78
4     24    20    10          0.63
5     24    20    10          0.44
6     24    20    10          0.52
7     12    20    30          0.65
8     12    20    30          0.67
9     12    20    30          0.86
10    12    12    10          0.30
11    12    12    10          0.45
12    12    12    10          0.26
13    12    20    10          0.49
14    12    20    10          0.58
15    12    20    10          0.47
16    12    12    30          0.48
17    12    12    30          0.56
18    12    12    30          0.35
19    24    12    10          0.53
20    24    12    10          0.74
21    24    12    10          0.63
22    24    20    30          0.85
23    24    20    30          0.98
24    24    20    30          0.81
25    11    16    20          0.36
26    11    16    20          0.47
27    11    16    20          0.55
28    18    11    20          0.53
29    18    11    20          0.74
30    18    11    20          0.60
31    18    16    8           0.58
32    18    16    8           0.35
33    18    16    8           0.25
34    25    16    20          0.77
35    25    16    20          0.93
```

```
36    25    16    20          0.81
37    18    21    20          0.62
38    18    21    20          0.93
39    18    21    20          0.68
40    18    16    32          0.86
41    18    16    32          0.94
42    18    16    32          0.96
43    18    16    20          0.75
43    18    16    20          0.73
45    18    16    20          0.62
;
proc glm data=info;
model Probability= Size Density Velocity Size*Density Size*Velocity
Density*Velocity Size*Size Density*Density Velocity*Velocity;
proc rsreg data=info;
model Probability= Size Density Velocity/LACKFIT;
run;
quit;
```

### SAS Output (Part B. Between-Subjects Raw Score)

```
Example 39B: Orthogonal, Between-Subjects-Subjects, Central-Composite Design (Raw Score)       1

The GLM Procedure

Number of Observations Read          45
Number of Observations Used          45

Dependent Variable: Probability

                                   Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F

Model                        9      1.25196924      0.13910769      11.05    <.0001

Error                       35      0.44075521      0.01259301

Corrected Total             44      1.69272444


R-Square     Coeff Var      Root MSE    Probability Mean

0.739618      17.66912      0.112219            0.635111


Source                      DF       Type I SS     Mean Square    F Value    Pr > F

Size                         1      0.40926848      0.40926848      32.50    <.0001
Density                      1      0.09707865      0.09707865       7.71    0.0088
Velocity                     1      0.58400600      0.58400600      46.38    <.0001
Size*Density                 1      0.06933750      0.06933750       5.51    0.0247
Size*Velocity                1      0.00770417      0.00770417       0.61    0.4394
Density*Velocity             1      0.03450417      0.03450417       2.74    0.1068
Size*Size                    1      0.02820012      0.02820012       2.24    0.1435
Density*Density              1      0.00393044      0.00393044       0.31    0.5799
Velocity*Velocity            1      0.01793972      0.01793972       1.42    0.2407
```

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Size** | **1** | **0.08385942** | **0.08385942** | **6.66** | **0.0142** |
| Density | 1 | 0.01498329 | 0.01498329 | 1.19 | 0.2828 |
| Velocity | 1 | 0.00264703 | 0.00264703 | 0.21 | 0.6494 |
| **Size*Density** | **1** | **0.06933750** | **0.06933750** | **5.51** | **0.0247** |
| Size*Velocity | 1 | 0.00770417 | 0.00770417 | 0.61 | 0.4394 |
| Density*Velocity | 1 | 0.03450417 | 0.03450417 | 2.74 | 0.1068 |
| Size*Size | 1 | 0.02552432 | 0.02552432 | 2.03 | 0.1634 |
| Density*Density | 1 | 0.00420672 | 0.00420672 | 0.33 | 0.5670 |
| Velocity*Velocity | 1 | 0.01793972 | 0.01793972 | 1.42 | 0.2407 |

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | -1.189941916 | 0.65442552 | -1.82 | 0.0776 |
| Size | 0.094616417 | 0.03666530 | 2.58 | 0.0142 |
| Density | 0.069643079 | 0.06384680 | 1.09 | 0.2828 |
| Velocity | 0.007852595 | 0.01712766 | 0.46 | 0.6494 |
| Size*Density | -0.002239583 | 0.00095444 | -2.35 | 0.0247 |

Dependent Variable: Probability

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Size*Velocity | 0.000298611 | 0.00038178 | 0.78 | 0.4394 |
| Density*Velocity | 0.000947917 | 0.00057266 | 1.66 | 0.1068 |
| Size*Size | -0.001276546 | 0.00089665 | -1.42 | 0.1634 |
| Density*Density | -0.001087680 | 0.00188189 | -0.58 | 0.5670 |
| Velocity*Velocity | -0.000375451 | 0.00031456 | -1.19 | 0.2407 |

The RSREG Procedure

Coding Coefficients for the Independent Variables

| Factor | Subtracted off | Divided by |
|---|---|---|
| Size | 18.000000 | 7.000000 |
| Density | 16.000000 | 5.000000 |
| Velocity | 20.000000 | 12.000000 |

Response Surface for Variable Probability

| | |
|---|---|
| Response Mean | 0.635111 |
| Root MSE | 0.112219 |
| **R-Square** | **0.7396** |
| Coefficient of Variation | 17.6691 |

| Regression | DF | Type I Sum of Squares | R-Square | F Value | Pr > F |
|---|---|---|---|---|---|

| Non-Professional | Acceptability of Photograph | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Photographer | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | Median |
| 1 | 5 | 2 | 3 | 4 | 1 | 2 | 6 | 2 | 4 | 3 | 1 | 5 | 2 | 1 | 1 | 3 | 4 | 2 | 2 | 2 | 5 | 3 | 1 | 2 | 1 | 2 |
| 2 | 6 | 4 | 3 | 5 | 2 | 6 | 7 | 3 | 4 | 5 | 3 | 7 | 5 | 4 | 1 | 3 | 6 | 2 | 4 | 6 | 5 | 6 | 7 | 4 | 5 | 5 |
| 3 | 1 | 3 | 4 | 6 | 2 | 7 | 5 | 2 | 4 | 1 | 4 | 3 | 6 | 5 | 2 | 4 | 1 | 4 | 3 | 2 | 5 | 6 | 4 | 1 | 3 | 4 |
| 4 | 7 | 5 | 7 | 7 | 6 | 2 | 4 | 7 | 3 | 4 | 7 | 7 | 2 | 7 | 6 | 7 | 2 | 1 | 7 | 7 | 6 | 7 | 7 | 5 | 7 | 7 |
| 5 | 1 | 3 | 4 | 5 | 7 | 5 | 6 | 2 | 3 | 7 | 6 | 4 | 6 | 4 | 6 | 7 | 5 | 6 | 7 | 5 | 6 | 6 | 7 | 7 | 3 | 6 |
| 6 | 3 | 2 | 1 | 4 | 2 | 5 | 3 | 1 | 7 | 2 | 4 | 1 | 2 | 3 | 2 | 5 | 1 | 3 | 2 | 4 | 2 | 1 | 3 | 5 | 2 | 2 |
| 7 | 5 | 6 | 6 | 1 | 1 | 2 | 6 | 6 | 7 | 6 | 3 | 7 | 5 | 6 | 7 | 7 | 6 | 2 | 3 | 6 | 7 | 7 | 1 | 5 | 6 | 6 |
| 8 | 1 | 5 | 4 | 1 | 7 | 5 | 2 | 6 | 4 | 5 | 3 | 1 | 5 | 4 | 7 | 3 | 5 | 4 | 6 | 2 | 5 | 4 | 6 | 5 | 7 | 5 |
| 9 | 3 | 1 | 4 | 5 | 6 | 2 | 1 | 0 | 5 | 4 | 2 | 7 | 3 | 5 | 4 | 3 | 1 | 4 | 5 | 6 | 4 | 3 | 4 | 5 | 1 | 4 |
| 10 | 4 | 3 | 1 | 5 | 3 | 3 | 7 | 1 | 4 | 2 | 3 | 5 | 7 | 3 | 4 | 5 | 1 | 3 | 3 | 1 | 2 | 4 | 5 | 3 | 3 | 3 |
| 11 | 3 | 7 | 4 | 6 | 7 | 7 | 1 | 7 | 7 | 6 | 7 | 2 | 4 | 7 | 6 | 7 | 5 | 7 | 2 | 4 | 7 | 3 | 7 | 7 | 7 | 7 |
| 12 | 2 | 3 | 6 | 4 | 5 | 5 | 1 | 2 | 7 | 5 | 4 | 6 | 7 | 4 | 5 | 1 | 5 | 6 | 3 | 4 | 2 | 4 | 5 | 7 | 5 | 5 |
| 13 | 2 | 1 | 3 | 1 | 4 | 1 | 6 | 3 | 2 | 1 | 1 | 1 | 5 | 2 | 1 | 1 | 4 | 2 | 1 | 1 | 6 | 2 | 1 | 1 | 1 | 1 |
| 14 | 1 | 5 | 4 | 7 | 5 | 3 | 4 | 7 | 2 | 6 | 7 | 3 | 4 | 3 | 5 | 7 | 6 | 4 | 3 | 5 | 6 | 2 | 4 | 7 | 5 | 5 |
| 15 | 5 | 1 | 7 | 3 | 2 | 4 | 5 | 1 | 1 | 2 | 4 | 7 | 3 | 5 | 1 | 4 | 2 | 4 | 4 | 3 | 4 | 7 | 5 | 1 | 5 | 4 |
| 16 | 7 | 6 | 4 | 6 | 5 | 2 | 7 | 6 | 3 | 4 | 7 | 6 | 7 | 5 | 6 | 2 | 6 | 5 | 3 | 7 | 6 | 2 | 4 | 6 | 5 | 6 |
| 17 | 5 | 2 | 3 | 2 | 4 | 5 | 3 | 2 | 1 | 2 | 2 | 7 | 5 | 3 | 3 | 1 | 4 | 5 | 1 | 2 | 2 | 4 | 3 | 3 | 2 | 3 |
| 18 | 6 | 1 | 3 | 5 | 7 | 4 | 5 | 2 | 4 | 5 | 5 | 6 | 4 | 2 | 4 | 7 | 5 | 3 | 7 | 4 | 5 | 3 | 6 | 2 | 5 | 5 |
| 19 | 1 | 3 | 6 | 7 | 2 | 4 | 6 | 4 | 2 | 6 | 7 | 3 | 3 | 5 | 6 | 7 | 7 | 6 | 4 | 2 | 7 | 6 | 7 | 2 | 6 | 6 |
| 20 | 2 | 5 | 7 | 3 | 2 | 3 | 5 | 1 | 2 | 6 | 4 | 5 | 4 | 7 | 5 | 2 | 6 | 5 | 1 | 4 | 5 | 5 | 3 | 6 | 7 | 5 |
| 21 | 4 | 5 | 2 | 1 | 3 | 6 | 1 | 1 | 2 | 4 | 2 | 3 | 1 | 5 | 2 | 3 | 5 | 2 | 2 | 3 | 1 | 6 | 2 | 1 | 1 | 2 |
| 22 | 6 | 5 | 4 | 7 | 6 | 1 | 7 | 5 | 6 | 7 | 5 | 4 | 6 | 1 | 4 | 3 | 6 | 7 | 7 | 4 | 7 | 5 | 6 | 5 | 6 | 6 |
| 23 | 1 | 3 | 7 | 5 | 6 | 7 | 2 | 7 | 7 | 3 | 7 | 6 | 7 | 5 | 4 | 7 | 7 | 3 | 7 | 7 | 4 | 7 | 7 | 6 | 7 | 7 |
| 24 | 6 | 1 | 4 | 2 | 1 | 5 | 4 | 3 | 5 | 4 | 1 | 7 | 4 | 5 | 3 | 1 | 4 | 6 | 2 | 3 | 5 | 4 | 1 | 7 | 4 | 4 |
| 25 | 4 | 6 | 7 | 3 | 1 | 4 | 2 | 7 | 5 | 2 | 5 | 4 | 6 | 7 | 6 | 5 | 3 | 1 | 6 | 4 | 2 | 3 | 5 | 6 | 7 | 5 |
| 26 | 3 | 3 | 7 | 2 | 4 | 5 | 6 | 7 | 6 | 2 | 5 | 7 | 6 | 7 | 6 | 4 | 7 | 3 | 3 | 5 | 6 | 7 | 6 | 5 | 7 | 6 |
| 27 | 7 | 3 | 4 | 3 | 5 | 7 | 5 | 6 | 7 | 7 | 6 | 7 | 4 | 7 | 3 | 7 | 7 | 4 | 7 | 6 | 7 | 5 | 7 | 7 | 7 | 7 |
| 28 | 4 | 5 | 1 | 3 | 5 | 7 | 3 | 1 | 4 | 2 | 6 | 5 | 1 | 7 | 5 | 3 | 6 | 5 | 2 | 4 | 5 | 7 | 6 | 5 | 3 | 5 |
| 29 | 7 | 6 | 2 | 1 | 5 | 7 | 6 | 5 | 2 | 3 | 7 | 6 | 5 | 7 | 6 | 6 | 6 | 6 | 5 | 2 | 7 | 3 | 7 | 4 | 4 | 6 |
| 30 | 5 | 4 | 1 | 2 | 3 | 5 | 4 | 6 | 4 | 1 | 2 | 3 | 4 | 5 | 7 | 6 | 3 | 1 | 2 | 4 | 3 | 6 | 4 | 1 | 4 | 4 |

**SAS Input**

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title 'Example 11: Kolmogorov-Smirnov Test';
data photos;
input Group $ rating count;
lines;
P 1 9
P 2 6
P 3 1
P 4 2
P 5 4
P 6 2
P 7 1
N 1 1
N 2 3
N 3 2
N 4 5
N 5 8
N 6 7
N 7 4
;
proc npar1way data=photos edf;
class Group;
var rating;
freq count;
quit;
```

| | | | | | |
|---|---|---|---|---|---|
| Linear | 3 | 1.090353 | 0.6441 | 28.86 | <.0001 |
| Quadratic | 3 | 0.050070 | 0.0296 | 1.33 | 0.2817 |
| Crossproduct | 3 | 0.111546 | 0.0659 | 2.95 | 0.0459 |
| Total Model | 9 | 1.251969 | 0.7396 | 11.05 | <.0001 |

| Residual | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Lack of Fit** | **5** | **0.119289** | **0.023858** | **2.23** | **0.0775** |
| **Pure Error** | **30** | **0.321467** | **0.010716** | | |
| **Total Error** | **35** | **0.440755** | **0.012593** | | |

| Parameter | DF | Estimate | Standard Error | t Value | Pr > \|t\| | Parameter Estimate from Coded Data |
|---|---|---|---|---|---|---|
| **Intercept** | **1** | **-1.189942** | 0.654426 | -1.82 | 0.0776 | 0.708101 |
| **Size** | **1** | **0.094616** | 0.036665 | 2.58 | 0.0142 | 0.131598 |
| **Density** | **1** | **0.069643** | 0.063847 | 1.09 | 0.2828 | 0.067416 |
| **Velocity** | **1** | **0.007853** | 0.017128 | 0.46 | 0.6494 | 0.160515 |
| **Size*Size** | **1** | **-0.001277** | 0.000897 | -1.42 | 0.1634 | -0.062551 |
| **Density*Size** | **1** | **-0.002240** | 0.000954 | -2.35 | 0.0247 | -0.078385 |
| **Density*Density** | **1** | **-0.001088** | 0.001882 | -0.58 | 0.5670 | -0.027192 |
| **Velocity*Size** | **1** | **0.000299** | 0.000382 | 0.78 | 0.4394 | 0.025083 |
| **Velocity*Density** | **1** | **0.000948** | 0.000573 | 1.66 | 0.1068 | 0.056875 |
| **Velocity*Velocity** | **1** | **-0.000375** | 0.000315 | -1.19 | 0.2407 | -0.054065 |

| Factor | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Size | 4 | 0.511834 | 0.127959 | 10.16 | <.0001 |
| Density | 4 | 0.205127 | 0.051282 | 4.07 | 0.0082 |
| Velocity | 4 | 0.644154 | 0.161039 | 12.79 | <.0001 |

Canonical Analysis of Response Surface Based on Coded Data

| Factor | Critical Value Coded | Uncoded |
|---|---|---|
| Size | 4.332517 | 48.327621 |
| Density | -5.337505 | -10.687523 |
| Velocity | -0.317968 | 16.184378 |

Predicted value at stationary point: 0.787740

| Eigenvalues | Eigenvectors Size | Density | Velocity |
|---|---|---|---|
| 0.003890 | -0.436292 | 0.841545 | 0.318515 |
| -0.045069 | 0.586540 | -0.002457 | 0.809917 |
| -0.102630 | 0.682364 | 0.540182 | -0.492528 |

Stationary point is a saddle point.

### Output Explanation (Part B. Between-Subjects Raw Scores)

By using raw scores of levels in the polynomial regression, the complete, second-order, empirical model that predicts the probability of target detection (P) as a function of the three display variables is: P = -1.1899 + 0.0946(Size) + 0.0696(Density) + 0.00079(Velocity) – 0.0022(Size x Density) + 0.0003(Size x Velocity) + 0.0009(Density x Velocity) – 0.0013($Size^2$) - 0.0011($Density^2$) - 0.0004($Velocity^2$). The $R^2$ value (0.74) indicates that approximately 74% of the variation in probability of target detection is accounted for by the second-order empirical model. The p-value (<0.001) is less than the specified significance level (0.05). Therefore, the relationship describing the probability is statistically significant. The predictor of target size is statistically significant (0.014) at the 0.05 level, and the linear-by-linear predictor of Size*Density is also significant (0.024). The models had equal levels of significance (<0.0001) and $R^2$ values (0.74). However, the regression models have different parameter estimates since the raw score values do not result in orthogonal partial regression weights. Note the differences and similarities between the coded values (Part A) and raw score (Part B) results. Shown below is a partially revised ANOVA summary table that uses the information provided by SAS. The complete ANOVA summary table for this design can be found in the Williges (2006) reference.

**CCD Revised ANOVA Summary Table (Orthogonal Raw Score Between-Subjects Design)**

| Source | DF | Type III Sum of Squares | F Value | Pr > F |
|---|---|---|---|---|
| Model | (9) | (1.25196924) | (11.05) | (<.0001) |
| Size | 1 | 0.08385942 | 6.66 | 0.0142 |
| Density | 1 | 0.01498329 | 1.19 | 0.2828 |
| Velocity | 1 | 0.00264703 | 0.21 | 0.6494 |
| Size*Density | 1 | 0.06933750 | 5.51 | 0.0247 |
| Size*Velocity | 1 | 0.00770417 | 0.61 | 0.4394 |
| Density*Velocity | 1 | 0.03450417 | 2.74 | 0.1068 |
| Size*Size | 1 | 0.02552432 | 2.03 | 0.1634 |
| Density*Density | 1 | 0.00420672 | 0.33 | 0.5670 |
| Velocity*Velocity | 1 | 0.01793972 | 1.42 | 0.2407 |
| Total Error | (35) | (0.440755) | | |
| Lack of Fit | 5 | 0.119289 | 2.23 | 0.0775 |
| Pure Error | 30 | 0.321467 | | |
| Corrected Total | 44 | 1.69272444 | | |

## Example 40: Blocked, Within-Subjects, Central-Composite Design

**(Click in this red rectangle to see the Reference Notes on Example 40.)**

### Problem

Location in Williges (2006) Table of Contents
Section 5, Topic 23. Central-Composite Designs (CCD), Part 23.4.2. Within-Subjects Example

Page(s) in Williges (2006) Reference Material: 802-805

Problem Description
A computer-generated Army surveillance display is tested to predict the effects of three target characteristics on the probability of target detection. The three parameters of interest are target size, target density, and target velocity. Three soldiers were tested in a within-subjects, central-composite design that was blocked across three testing days. Is the complete second-order empirical model significant ($p < 0.05$)? Which predictors are significant and do significant higher-order predictors exist ($p < 0.05$)?

Context/Purpose
Develop a complete second-order empirical model that predicts the probability of target detection as a function of target size, target density, and target velocity.

Statistical Decision Criteria
Use an orthogonal blocked, second-order, within-subjects, central-composite design to develop the polynomial regression model and conduct an ANOVA on regression to test for significance at the 0.05 level. The coded value of $\alpha$ is set at ±1.871 to keep the effect of testing days orthogonal to the second-order empirical model.

### SAS Input

**(Click in this blue rectangle to open the following SAS Input directly in the SAS Editor.)**

```
options nodate nocenter pageno=1;
title'Example 40: Blocked, Within-Subjects, Central-Composite Design
(Coded)';
data info;
input Treatment Block Subject Size Density Velocity Probability;
lines;
1     1     1      1          -1           1          0.70
1     1     2      1          -1           1          0.82
1     1     3      1          -1           1          0.78
2     1     1      1           1          -1          0.63
2     1     2      1           1          -1          0.44
2     1     3      1           1          -1          0.52
3     1     1     -1           1           1          0.65
3     1     2     -1           1           1          0.67
3     1     3     -1           1           1          0.86
4     1     1     -1          -1          -1          0.30
4     1     2     -1          -1          -1          0.45
4     1     3     -1          -1          -1          0.26
```

```
5     2     1     -1          1          -1          0.49
5     2     2     -1          1          -1          0.58
5     2     3     -1          1          -1          0.47
6     2     1     -1         -1           1          0.48
6     2     2     -1         -1           1          0.56
6     2     3     -1         -1           1          0.35
7     2     1      1         -1          -1          0.53
7     2     2      1         -1          -1          0.74
7     2     3      1         -1          -1          0.63
8     2     1      1          1           1          0.85
8     2     2      1          1           1          0.98
8     2     3      1          1           1          0.81
9     3     1     -1.871      0           0          0.36
9     3     2     -1.871      0           0          0.47
9     3     3     -1.871      0           0          0.55
10    3     1      0         -1.871       0          0.53
10    3     2      0         -1.871       0          0.74
10    3     3      0         -1.871       0          0.60
11    3     1      0          0          -1.871      0.58
11    3     2      0          0          -1.871      0.35
11    3     3      0          0          -1.871      0.25
12    3     1      1.871      0           0          0.77
12    3     2      1.871      0           0          0.93
12    3     3      1.871      0           0          0.81
13    3     1      0          1.871       0          0.62
13    3     2      0          1.871       0          0.93
13    3     3      0          1.871       0          0.68
14    3     1      0          0           1.871      0.86
14    3     2      0          0           1.871      0.94
14    3     3      0          0           1.871      0.96
15    3     1      0          0           0          0.75
15    3     2      0          0           0          0.73
15    3     3      0          0         0            0.62
;
proc glm data=info;
model Probability= Size Density Velocity Size*Density Size*Velocity
Density*Velocity Size*Size Density*Density Velocity*Velocity;
proc rsreg data=info;
model Probability= Size Density Velocity/LACKFIT;
run;
proc glm data=info;
class Block Subject;
model Probability= Block Subject;
run;
quit;
```

### SAS Output

```
Example 40: Blocked, Within-Subjects, Central-Composite Design (Coded)                    1

The GLM Procedure

Number of Observations Read          45
Number of Observations Used          45
```

The GLM Procedure
Dependent Variable: Probability

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | **9** | **1.29927815** | **0.14436424** | **12.84** | **<.0001** |
| Error | 35 | 0.39344630 | 0.01124132 | | |
| Corrected Total | 44 | 1.69272444 | | | |

| R-Square | Coeff Var | Root MSE | Probability Mean |
|---|---|---|---|
| 0.767566 | 16.69395 | 0.106025 | 0.635111 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Size | 1 | 0.43493641 | 0.43493641 | 38.69 | <.0001 |
| Density | 1 | 0.09098767 | 0.09098767 | 8.09 | 0.0074 |
| Velocity | 1 | 0.65424252 | 0.65424252 | 58.20 | <.0001 |
| Size*Density | 1 | 0.06933750 | 0.06933750 | 6.17 | 0.0179 |
| Size*Velocity | 1 | 0.00770417 | 0.00770417 | 0.69 | 0.4134 |
| Density*Velocity | 1 | 0.03450417 | 0.03450417 | 3.07 | 0.0885 |
| Size*Size | 1 | 0.00327847 | 0.00327847 | 0.29 | 0.5926 |
| Density*Density | 1 | 0.00053902 | 0.00053902 | 0.05 | 0.8279 |
| Velocity*Velocity | 1 | 0.00374823 | 0.00374823 | 0.33 | 0.5673 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Size** | **1** | **0.43493641** | **0.43493641** | **38.69** | **<.0001** |
| **Density** | **1** | **0.09098767** | **0.09098767** | **8.09** | **0.0074** |
| **Velocity** | **1** | **0.65424252** | **0.65424252** | **58.20** | **<.0001** |
| **Size*Density** | **1** | **0.06933750** | **0.06933750** | **6.17** | **0.0179** |
| Size*Velocity | 1 | 0.00770417 | 0.00770417 | 0.69 | 0.4134 |
| Density*Velocity | 1 | 0.03450417 | 0.03450417 | 3.07 | 0.0885 |
| Size*Size | 1 | 0.00533015 | 0.00533015 | 0.47 | 0.4956 |
| Density*Density | 1 | 0.00055274 | 0.00055274 | 0.05 | 0.8258 |
| Velocity*Velocity | 1 | 0.00374823 | 0.00374823 | 0.33 | 0.5673 |

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| **Intercept** | **0.6669765192** | 0.05883015 | 11.34 | <.0001 |
| **Size** | **0.0983078202** | 0.01580461 | 6.22 | <.0001 |
| **Density** | **0.0449641571** | 0.01580461 | 2.85 | 0.0074 |
| **Velocity** | **0.1205714729** | 0.01580461 | 7.63 | <.0001 |
| **Size*Density** | **-.0537500000** | 0.02164228 | -2.48 | 0.0179 |

Dependent Variable: Probability

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| **Size*Velocity** | **0.0179166667** | 0.02164228 | 0.83 | 0.4134 |
| **Density*Velocity** | **0.0379166667** | 0.02164228 | 1.75 | 0.0885 |
| **Size*Size** | **-.0147471234** | 0.02141638 | -0.69 | 0.4956 |
| **Density*Density** | **-.0047489545** | 0.02141638 | -0.22 | 0.8258 |
| **Velocity*Velocity** | **-.0123666070** | 0.02141638 | -0.58 | 0.5673 |

The RSREG Procedure

Coding Coefficients for the Independent Variables

| Factor | Subtracted off | Divided by |
|---|---|---|
| Size | 0 | 1.871000 |
| Density | 0 | 1.871000 |
| Velocity | 0 | 1.871000 |

Response Surface for Variable Probability

| | |
|---|---|
| Response Mean | 0.635111 |
| Root MSE | 0.106025 |
| R-Square | 0.7676 |
| Coefficient of Variation | 16.6939 |

| Regression | DF | Type I Sum of Squares | R-Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Linear | 3 | 1.180167 | 0.6972 | 34.99 | <.0001 |
| Quadratic | 3 | 0.007566 | 0.0045 | 0.22 | 0.8788 |
| Crossproduct | 3 | 0.111546 | 0.0659 | 3.31 | 0.0313 |
| Total Model | 9 | 1.299278 | 0.7676 | 12.84 | <.0001 |

| Residual | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Lack of Fit | 5 | 0.071980 | 0.014396 | 1.34 | 0.2733 |
| Pure Error | 30 | 0.321467 | 0.010716 | | |
| Total Error | 35 | 0.393446 | 0.011241 | | |

| Parameter | DF | Estimate | Standard Error | t Value | Pr > \|t\| | Parameter Estimate from Coded Data |
|---|---|---|---|---|---|---|
| Intercept | 1 | 0.666977 | 0.058830 | 11.34 | <.0001 | 0.666977 |
| Size | 1 | 0.098308 | 0.015805 | 6.22 | <.0001 | 0.183934 |
| Density | 1 | 0.044964 | 0.015805 | 2.85 | 0.0074 | 0.084128 |
| Velocity | 1 | 0.120571 | 0.015805 | 7.63 | <.0001 | 0.225589 |
| Size*Size | 1 | -0.014747 | 0.021416 | -0.69 | 0.4956 | -0.051624 |
| Density*Size | 1 | -0.053750 | 0.021642 | -2.48 | 0.0179 | -0.188159 |
| Density*Density | 1 | -0.004749 | 0.021416 | -0.22 | 0.8258 | -0.016624 |
| Velocity*Size | 1 | 0.017917 | 0.021642 | 0.83 | 0.4134 | 0.062720 |
| Velocity*Density | 1 | 0.037917 | 0.021642 | 1.75 | 0.0885 | 0.132733 |
| Velocity*Velocity | 1 | -0.012367 | 0.021416 | -0.58 | 0.5673 | -0.043291 |

| Factor | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Size | 4 | 0.517308 | 0.129327 | 11.50 | <.0001 |
| Density | 4 | 0.195382 | 0.048846 | 4.35 | 0.0059 |
| Velocity | 4 | 0.700199 | 0.175050 | 15.57 | <.0001 |

Canonical Analysis of Response Surface Based on Coded Data

| Factor | Critical Value Coded | Uncoded |
|---|---|---|
| Size | 5.390717 | 10.086031 |
| Density | 0.387870 | 0.725706 |
| Velocity | 7.105130 | 13.293698 |

Predicted value at stationary point: 1.980480

| Eigenvalues | Eigenvectors Size | Density | Velocity |
|---|---|---|---|
| 0.071875 | -0.523661 | 0.792020 | 0.313822 |
| -0.016517 | 0.575447 | 0.057211 | 0.815836 |
| -0.166897 | 0.628204 | 0.607809 | -0.485724 |

Stationary point is a saddle point.

```
The GLM Procedure

   Class Level Information

Class         Levels    Values

Block              3    1 2 3

Subject            3    1 2 3


Number of Observations Read          45
Number of Observations Used          45

Dependent Variable: Probability

                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F

Model                        4      0.11379341      0.02844835       0.72    0.5829

Error                       40      1.57893103      0.03947328

Corrected Total             44      1.69272444


R-Square     Coeff Var      Root MSE    Probability Mean

0.067225      31.28253      0.198679            0.635111


Source                      DF       Type I SS     Mean Square    F Value    Pr > F

Block                        2      0.04917563      0.02458782       0.62    0.5415
Subject                      2      0.06461778      0.03230889       0.82    0.4483


Source                      DF     Type III SS     Mean Square    F Value    Pr > F

Block                        2      0.04917563      0.02458782       0.62    0.5415
Subject                      2      0.06461778      0.03230889       0.82    0.4483
```

## Output Explanation

By using coded-values of levels in the polynomial regression, the complete, second-order, empirical model that predicts the probability of target detection (P) as a function of the three display variables is: P = 0.6670 + 0.0983(Size) + 0.0450(Density) + 0.1206(Velocity) – 0.0538(Size x Density) + 0.0179(Size x Velocity) + 0.0379(Density x Velocity) – 0.0147($Size^2$) - 0.0047($Density^2$) - 0.0124($Velocity^2$). The $R^2$ value (0.77) indicates that approximately 77% of the variation in probability of target detection is accounted for by the second-order empirical model. The p-value for the regression model (<0.001) is less than the specified significance level (0.05). Therefore, the relationship describing the probability of detection is statistically significant. However, the ANOVA results indicate that testing days is not significant (p = 0.54) nor is the effect due to the subjects (p = 0.45). The partial regression weights for target size (p <

0.001), density (p = 0.0074), velocity (p < 0.001), and the linear-by-linear interaction of size and density (p = 0.0179) all have a significant effect on the probability of target detection at the 0.05 level. Shown below is a partially revised ANOVA summary table that uses, the information provided by SAS. The complete ANOVA summary table for this design can be found in the Williges (2006) reference. Note that the SAS output has different lack of fit and error values because SAS requires two separate programs to obtain the blocking and within-subjects results. To obtain the correct calculation, see the Williges (2006) reference.

**CCD Summary Table (Blocked Coded Within-Subjects Design)**

| Source | DF | Type III Sum of Squares | F Value | Pr > F |
|---|---|---|---|---|
| Model | (9) | (1.29927815) | (12.84) | (<.0001) |
| Size | 1 | 0.43493641 | 38.69 | <.0001 |
| Density | 1 | 0.09098767 | 8.09 | 0.0074 |
| Velocity | 1 | 0.65424252 | 58.20 | <.0001 |
| Size*Density | 1 | 0.06933750 | 6.17 | 0.0179 |
| Size*Velocity | 1 | 0.00770417 | 0.69 | 0.4134 |
| Density*Velocity | 1 | 0.03450417 | 3.07 | 0.0885 |
| Size*Size | 1 | 0.00533015 | 0.47 | 0.4956 |
| Density*Density | 1 | 0.00055274 | 0.05 | 0.8258 |
| Velocity*Velocity | 1 | 0.00374823 | 0.33 | 0.5673 |
| Total Error | (35) | (0.393446) | | |
| Lack of Fit* | 3 | 0.071980 | 1.34 | 0.2733 |
| Block | 2 | 0.04917563 | 0.62 | 0.5415 |
| Subject | 2 | 0.06461778 | 0.82 | 0.4483 |
| Error* | 28 | 0.2568 | | |
| Corrected Total | 44 | 1.692724 | | |

*These values have been modified from the SAS output to use the Williges (2006) error term corrected for blocks and the subject effect that is not calculated in SAS.

## Acknowledgements

This research was conducted for the Human Research and Engineering Directorate (HRED) of the U.S. Army Research Laboratory through the Collaborative Technology Alliance on Advanced Decision Architectures (ADA) prime contract to Micro Analysis and Design. The authors acknowledge their financial support of this project. Michael Strub was the HRED manager of ADA, and Sam Middlebrooks directed the HRED formative evaluation of the draft report. This report was improved by several members of the HRED staff including James Ainsworth, Elizabeth Bowman, Otto Heuckeroth, Andrea Krausman, Sam Middlebrooks, Paul Rose, and Ron Spencer. We thank them for their valuable suggestions and reviews of the draft materials. We also greatly appreciate the help provided by Juwei Sheng, Raina Duan, and David Edwards of the Statistical Consulting Center at Virginia Tech for assistance with some of the SAS programs, Beverly Williges for proofreading and statistical consulting, and Jason Lockhart the Director of High Performance Computing at Virginia Tech for assistance in producing the computer media used for report delivery.

## References

Cody, R.P. and Smith, J.K. (1997). *Applied statistics and the SAS® programming language*. New Jersey: Prentice Hall.

SAS Institute (2004). SAS system version 9.1.3 for Windows and online documentation. North Carolina: The SAS Institute.

Siegel, S. and Castellan, N.J. (1988). *Nonparametric statistics for the behavioral sciences*. (2nd Ed.) New York: McGraw-Hill, Inc.

Williges, R.C. (2006). Human factors experimental design and analysis reference. Blacksburg, VA: Virginia Polytechnic Institute and State University, Technical Report HFEEC-06-01.

NO. OF COPIES ORGANIZATION

1 (PDF ONLY) DEFENSE TECHNICAL INFORMATION CTR DTIC OCA 8725 JOHN J KINGMAN RD STE 0944 FORT BELVOIR VA 22060-6218

1 (PDF ONLY) US ARMY RSRCH DEV & ENGRG CMD SYSTEMS OF SYSTEMS INTEGRATION AMSRD SS T 6000 6TH ST STE 100 FORT BELVOIR VA 22060-5608

1 (PDF ONLY) DIRECTOR US ARMY RESEARCH LAB IMNE ALC IMS 2800 POWDER MILL RD ADELPHI MD 20783-1197

1 (PDF ONLY) DIRECTOR US ARMY RESEARCH LAB AMSRD ARL CI OK TL 2800 POWDER MILL RD ADELPHI MD 20783-1197

1 (PDF ONLY) DIRECTOR US ARMY RESEARCH LAB AMSRD ARL CI OK T 2800 POWDER MILL RD ADELPHI MD 20783-1197

(PDF ONLY) ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR M DR M STRUB 6359 WALKER LANE SUITE 100 ALEXANDRIA VA 22310

(PDF ONLY) ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR ML J MARTIN MYER CENTER RM 2D311 FT MONMOUTH NJ 07703-5601

(PDF ONLY) ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MZ A DAVISON 199 E 4TH ST STE C TECH PARK BLDG 2 FT LEONARD WOOD MO 65473-1949

(PDF ONLY) ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MD T COOK BLDG 5400 RM C242 REDSTONE ARSENAL AL 35898-7290

NO. OF COPIES ORGANIZATION

(PDF ONLY) COMMANDANT USAADASCH ATTN AMSRD ARL HR ME A MARES 5800 CARTER RD FT BLISS TX 79916-3802

(PDF ONLY) ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MO J MINNINGER BLDG 5400 RM C242 REDSTONE ARSENAL AL 35898-7290

(PDF ONLY) ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MM V RICE-BERG BLDG 4011 RM 217 1750 GREELEY RD FT SAM HOUSTON TX 78234-5094

(PDF ONLY) ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MG R SPINE BUILDING 333 PICATINNY ARSENAL NJ 07806-5000

(PDF ONLY) ARL HRED ARMC FLD ELMT ATTN AMSRD ARL HR MH C BURNS BLDG 1467B ROOM 336 THIRD AVENUE FT KNOX KY 40121

(PDF ONLY) ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MJ D DURBIN AVNC FIELD ELEMENT BLDG 4506 (DCD) RM 107 FT RUCKER AL 36362-5000

(PDF ONLY) ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MK MR J REINHART 10125 KINGMAN RD FT BELVOIR VA 22060-5828

4 ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MV HQ USAOTC S MIDDLEBROOKS (PDF + 4 HARD COPIES) 91012 STATION AVE ROOM 348 FT HOOD TX 76544-5073

(PDF ONLY) ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MY M BARNES 2520 HEALY AVE STE 1172 BLDG 51005 FT HUACHUCA AZ 85613-7069

(PDF ONLY) ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MP D UNGVARSKY BATTLE CMD BATTLE LAB 415 SHERMAN AVE UNIT 3 FT LEAVENWORTH KS 66027-2326

NO. OF
COPIES ORGANIZATION

(PDF ONLY) ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MJK
J HANSBERGER
JFCOM JOINT EXPERIMENTATION J9
JOINT FUTURES LAB
115 LAKEVIEW PARKWAY SUITE B
SUFFOLK VA 23435

(PDF ONLY) ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MQ M R FLETCHER
US ARMY SBCCOM NATICK SOLDIER CTR
AMSRD NSC SS E BLDG 3 RM 341
NATICK MA 01760-5020

(PDF ONLY) ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MY DR J CHEN
12423 RESEARCH PARKWAY
ORLANDO FL 32826

(PDF ONLY) ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MS MR C MANASCO
SIGNAL TOWERS 118 MORAN HALL
FORT GORDON GA 30905-5233

(PDF ONLY) ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MU M SINGAPORE
6501 E 11 MILE RD MAIL STOP 284
BLDG 200A 2ND FL RM 2104
WARREN MI 48397-5000

(PDF ONLY) ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MF MR
C HERNANDEZ
BLDG 3040 RM 220
FORT SILL OK 73503-5600

(PDF ONLY) ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MW E REDDEN
BLDG 4 ROOM 332
FT BENNING GA 31905-5400

(PDF ONLY) ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MN R SPENCER
DCSFDI HF
HQ USASOC BLDG E2929
FORT BRAGG NC 28310-5000

(PDF ONLY) ARMY G1
ATTN DAPE MR B KNAPP
300 ARMY PENTAGON ROOM 2C489
WASHINGTON DC 20310-0300

NO. OF
COPIES ORGANIZATION

(PDF ONLY) ARL HRED LIAISON
PHYSICAL SCIENCES LAB
ATTN THOMAS M COOK
PO BOX 30002
LAS CRUCES NM 88003-8002

(PDF ONLY) HQ USASETAF
ATTN ASESE SA M GOLDEN
UNIT 31401
BOX 72
APO AE 09630

(PDF ONLY) VA POLYTECH INST & STATE UNIV
ATTN ROBT WILLIGES
3296 LAUREL DRIVE
BLACKSBURG VA 24060-1779

(PDF ONLY) VA POLYTECH INST & STATE UNIV
ATTN T SMITH-JACKSON
2001 BIRCHLEAF LANE
BLACKSBURG VA 24060

(PDF ONLY) VA POLYTECH INST & STATE UNIV
ATTN B KLEINER
1358 FALCON RIDGE RD
BLACKSBURG VA 24060

(PDF ONLY) VA POLYTECH INST & STATE UNIV
ATTN P KOELLING
3005 LANCASTER DRIVE
BLACKSBURG VA 24060

ABERDEEN PROVING GROUND

1 DIRECTOR
US ARMY RSCH LABORATORY
ATTN AMSRD ARL CI OK (TECH LIB)
BLDG 4600 (PDF + 1 HARD COPY)

1 DIRECTOR
US ARMY RSCH LABORATORY
ATTN AMSRD ARL CI OK TP S FOPPIANO
BLDG 459 (PDF + 1 HARD COPY)

(PDF ONLY) DIRECTOR
US ARMY RSCH LABORATORY
ATTN AMSRD ARL HR F PARAGALLO
BLDG 459

(PDF ONLY) DIRECTOR
US ARMY RSCH LABORATORY
ATTN AMSRD ARL HR M T HADUCH
BLDG 459

(PDF ONLY) DIRECTOR
US ARMY RSCH LABORATORY
ATTN AMSRD ARL HR MB J HAWLEY
BLDG 459