(12)

Technical Report 541

ADA 125 742

# RELIABILITY ESTIMATION FOR AGGREGATED DATA: APPLICATIONS FOR ORGANIZATIONAL RESEARCH

Roland J. Hart and Stephen C. Bradshaw

ARI FIELD UNIT AT PRESIDIO OF MONTEREY, CALIFORNIA

DTIC
ELECTE
JUN 24 1983
S D
E

U. S. Army
Research Institute for the Behavioral and Social Sciences

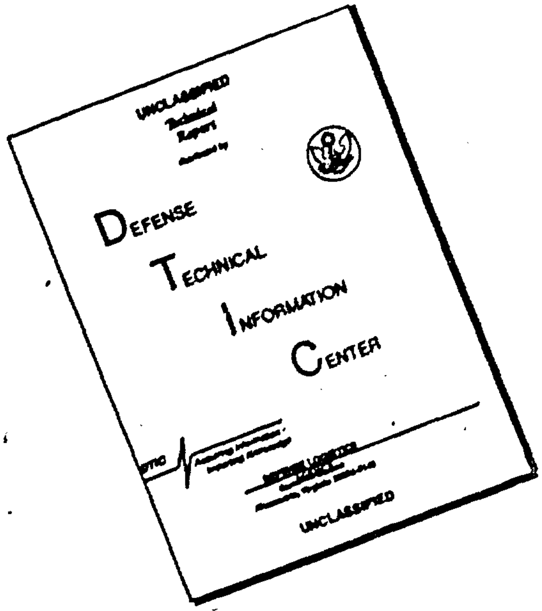
July 1981

Approved for public release, distribution unlimited.

DTIC FILE COPY

83 06 24 002

# DISCLAIMER NOTICE




**THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.**

# U. S. ARMY RESEARCH INSTITUTE

# FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel


JOSEPH ZEIDNER
Technical Director

L. NEALE COSBY
Colonel, IN
Commander


---


NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-TST, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: 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.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER<br>Technical Report 541 | 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle)<br>RELIABILITY ESTIMATION FOR AGGREGATED DATA: APPLICATIONS FOR ORGANIZATIONAL RESEARCH | | 5. TYPE OF REPORT & PERIOD COVERED<br>-- |
| | | 6. PERFORMING ORG. REPORT NUMBER<br>-- |
| 7. AUTHOR(s)<br>Roland J. Hart and Stephen C. Bradshaw | | 8. CONTRACT OR GRANT NUMBER(s)<br>-- |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>US Army Research Institute for the Behavioral and Social Sciences Field Unit, PO Box 5787, Presidio of Monterey, CA 93940 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS<br>2Q162717A766 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>US Army Research Institute for the Behavioral and Social Sciences<br>5001 Eisenhower Ave., Alexandria, VA 22333 | | 12. REPORT DATE<br>July 1981 |
| | | 13. NUMBER OF PAGES<br>34 |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)<br>-- | | 15. SECURITY CLASS. (of this report)<br>Unclassified |
| | | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE --  |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

--

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Measurement Error
Reliability Estimation
Organizational Research
Aggregated Data

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In order to study organizations it is important to be able to measure organizational functioning with a minimum of error. The report that follows provides the statistical tools necessary to measure the extent of error that exists in survey data, and organizational record data. Traditional methods of measuring error are either inappropriate or incomplete when applied to organizational groups, necessitating the statistical development given here. Appropriate methods of measuring error are particularly important when organizational change
(continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract (continued)

is being studied. In this case, the same variables are measured at more than one point in time. The investigator wants to identify real organizational change. However, real change cannot be separated from changes in measurement error, unless separate estimates of measurement error are available at each point in time. This paper tells how to get separate error estimates so that real organizational change can be studied.

When studying groups within organizations, what level of the hierarchy should be studied? A statistical technique for estimating the level of the hierarchy that actually controls the subject matter at hand is provided. This measure can be used as a guide for selecting groups at appropriate levels of hierarchy for study.

These statistical techniques provide improved procedures for studying the operation of the Army and other organizations.

Technical Report 541

# RELIABILITY ESTIMATION FOR AGGREGATED DATA: APPLICATIONS FOR ORGANIZATIONAL RESEARCH

Roland J. Hart and Stephen C. Bradshaw

Accession For
NTIS GRA&I X
DTIC TAB
Unannounced
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or Special
A

Submitted by:
James A. Thomas, Chief
ARI FIELD UNIT AT PRESIDIO OF MONTEREY, CALIFORNIA

Approved by:
E. Ralph Dusek, Director
TRAINING RESEARCH LABORATORY

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333

Office, Deputy Chief of Staff for Personnel
Department of the Army

July 1981

Army Project Number
2Q162717A766

Racial Harmony and Unit Effectiveness

Approved for public release; distribution unlimited.

ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

## FOREWORD

Leaders in any complex organization like the Army are constantly required to make decisions intended to improve organizational performance. Effective analysis and decision making by leaders require an understanding of organizational functioning and the dynamics of organizational change in theory and practice. Research can be designed to assist leaders in better understanding how their organization functions and how they may be improved. However, for such research to provide sound guidance to leaders, the methods that are employed must be capable of handling the complexities of dynamic individual and group interaction. Unfortunately, many of the methods currently employed by social scientists are best suited to handling less complex forms of data.

The purpose of this report is to provide researchers with statistical tools that will assist them in analyzing complex forms of data. The focus of this report is on techniques for estimating measurement error, using scores that are aggregated by group. These scores are useful for evaluating group dynamics in organizations as complex as the Army.

JOSEPH ZEIDNER
Technical Director

## RELIABILITY ESTIMATION FOR AGGREGATED DATA: APPLICATIONS FOR ORGANIZATIONAL RESEARCH

### BRIEF

Requirement:

In order to study organizations it is important to be able to measure organizational functioning with a minimum of error. The report that follows provides the statistical tools necessary to measure the extent of error that exists in survey data, and organizational record data. Traditional methods of measuring error are either inappropriate or incomplete when applied to organizational groups, necessitating the statistical development given here. Appropriate methods of measuring error are particularly important when organizational change is being studied. In this case, the same variables are measured at more than one point in time. The investigator wants to identify real organizational change. However, real change cannot be separated from changes in measurement error, unless separate estimates of measurement error are available at each point in time. This paper tells how to get separate error estimates so that real organizational change can be studied.

Procedure:

When research is conducted in an organizational setting, group units of analysis are often required. When group units of analysis are used, the values of the variables generally consist of mean scores that have been aggregated across both survey items and respondents within groups. Analysis of variance was used here to derive the appropriate reliability formulas for these aggregated scores. From the definition of reliability, which involves the ratio of true to total variance, formulas are derived by finding the mean square components that are equivalent to the reliability definition. This requires use of expected mean squares for the unit of analysis term and other "error" terms. Since the aggregated scores typically contain repeated observations across items as well as survey respondents, with respondents nested within groups, a split-plot (repeated-measures) design can usually describe the structure of the data, with a hierarchical structure added also as needed. This split-plot design contains two "error" terms--a split-plot (within-subjects) error term typically associated with inter-item agreement, and a whole plot (between-subjects) error term associated with consensus between respondents. Both types of error can enter into the reliability formula for aggregated scores, depending on whether survey items and respondents are considered to be fixed or random, which in turn depends on the sampling plan. For example, respondents may be fixed (or partially fixed) if the populations of small groups are exhaustively sampled, or nearly so. When respondents are fixed, the appropriate reliability formula is not the same as when respondents are random.

Findings:

Most of the literature on organizations using group units of analysis, have estimated reliability either incorrectly or inconsistently.

The survey construction and item analysis techniques that typically maximize inter-item agreement, may tend to reduce consensus between respondents, so that surveys like the Survey of Organizations, that were initially constructed to maximize inter-item agreement, may have poor reliability when consensus between respondents is desired.

When studying groups within organizations, what level of the hierarchy should be studied? A statistical technique for estimating the level of the heirarchy that actually controls the subject matter at hand is provided. This measure can be used as a guide for selecting groups at appropriate levels of heirarchy for study.

Utilization of Findings:

These statistical techniques provide improved procedures for studying the operation of the Army and other organizations. These techniques are an essential prerequisite to more advanced time-series procedures that are needed to study organizational change. If an investigator wishes to examine real organizational change, the change must take into account changes in measurement error. Sometimes change appears to be real but is due solely to changes in measurement error. Change in measurement error instead of real change can be used as a plausible alternative explanation for almost any set of results involving organizational change. If separate estimates of measurement error are available at each point in time, measurement error can be taken into account. This paper provides the tools needed to get appropriate <u>internal consistency</u> estimates of measurement error, and to show how these estimates change with time. Once these estimates are found, real organizational change, as distinct from changes in measurement accuracy, can be pinpointed.

RELIABILITY ESTIMATION FOR AGGREGATED DATA: APPLICATIONS FOR ORGANIZATIONAL RESEARCH

CONTENTS

TABLES

# RELIABILITY ESTIMATION FOR AGGREGATED DATA: APPLICATIONS FOR ORGANIZATIONAL RESEARCH

With the growth of organizational development over the last twenty years there has been an increase in field research on the functioning of intact organizations (Porras, 1979). Such field research has obvious advantages over laboratory research in terms of the possibilities for external validity, but at the same time researchers working with intact organizations face a variety of methodological questions that have not been satisfactorily answered to date.

One very basic question involves the selection of the unit of analysis for the research design. Individuals are not the appropriate unit of analysis to test most hypotheses about group functioning. When individuals are not appropriate units, which of many possible groups, at what level of the organizational hierarchy should be selected? The answer will be suggested by the hypotheses and organizational structure. The researcher wishes to select units that are responsible for and have control over the dependent variables. While organizational structure and the hypotheses may suggest which groups at what hierarchical level control particular variables, and thus provide an appropriate unit of analysis, the researcher has no way to test this hypothesis to find out if in fact groups at one level of the hierarchy provide a better unit of analysis than groups at another level. In principle, if groups at one level of the hierarchy are responsible for and have control over particular dependent variables, then we should find homogeneity within and heterogeneity between the independently operating groups on the dependent measures (see Jones & Jones, 1975; Bass, Valenzis, Farrow, & Solomon, 1975). This phenomenon will be called the principle of synchronization, and will be used later to show how to select appropriate units of analysis.

Evidence that researchers in the field are having trouble selecting units of analysis is suggested by the inconsistency with which a particular unit of analysis is used. Once a given unit of analysis is selected, this same unit should be used for stating hypotheses, calculating reliabilities and norms (when survey feedback is involved), estimating validity, and generalizing to new populations. A common problem is for researchers to state hypotheses and generalizations in terms of intact organizational groups, but to calculate reliabilities and estimate validity using individuals (see Bowers, 1973; also Passmore, 1976, and Torbert, 1973 for a critique of inconsistent use of units of analysis). The researcher may estimate validity with groups but calculate reliabilities using individuals (see Taylor & Bowers, 1972, p. 54 for alternation between using groups and individuals in calculating reliabilities).

The researcher who tries to use units of analysis consistently by computing reliabilities on the appropriate group units, faces difficulties since an adequate outline of procedures for estimating reliability on aggregated scores does not exist. Survey responses are aggregated across both items and respondents within each group to produce the dependent variable scores. The sources of true and error variance differ in these aggregated scores from the same sources of variance in individual level scores, since the structure of the data differs in the two cases, and for this reason the formulas for estimating reliability on aggregated scores can differ from the common formulas used with individuals. Some researchers have looked at inter-item agreement, and others at agreement between respondents within groups, but none have examined both sources of agreement in an integrated way. Researchers have looked at inter-item agreement by computing, for example, Cronbach's alpha on either individuals or on data aggregated over the unit of analysis for each item (see Taylor & Bowers, 1972); and at agreement between respondents by using either a variation of the intraclass correlation (see Jones & Jones, 1977; Ebel, 1951; Bass et al., 1975) or an iterative jacknife procedure (Schneider, 1972; Schneider & Bartlett, 1970).

Estimates of construct validity (Cronbach and Meehl, 1955) are in many cases dependent upon adequate measures of the reliability of the variables involved. Construct validity consists of hypotheses that make up nomological networks of expected relationships. The expected relationships involve expectations about differential levels of association among variables. Differential levels of association are frequently studied using regression or path analyses, or cross-lagged correlation analysis (see Kenny, 1975). Statistics that measure degrees of association among variables are a function of the variables' reliability as well as the degree of association in the population (McNemar, 1969, p. 163). Any attempt to measure differential levels of association must control for differential levels of reliability, or demonstrate that differential levels of reliability don't exist (Kenny, 1975; Jöreskog & Sörbom, 1979, chap. 4). Failure to calculate reliabilities provides alternate explanations for any set of results. In this sense, it is not possible to establish construct validity without taking into account measurement error first, no matter what method of analysis is used--regression, path, or cross-lagged panel correlation. In this way estimation of validity is dependent on the measurement of reliability.

The purpose here, then, is (a) to provide criteria for selecting appropriate units of analysis within intact organizations, and (b) to provide the appropriate procedures for calculating internal consistency reliabilities on the aggregated group scores. These internal consistency reliabilities are especially important in studies of organizational change. They can be used to identify possible reliability shifts over time. Real organizational changes can then be separated from changes in measurement error.

An important advantage of using group units over the common approach of using individuals, is that it allows the researcher to study the nature of the social interaction that occurs between subgroups within the unit--between blacks and whites, superiors and subordinates, parents and children--in a way that is not possible when individuals alone are the unit (see Hart, 1978, to illustrate this application). This is an advantage that has not been recognized, even by researchers with appropriate group data (see Taylor & Bowers, 1972). The structure of the data that allows interaction to be studied will be illustrated.

## Analysis of Variance

Analysis of variance can be used both for reliability estimation (see Winer 1971, pp. 283-296; Myers, 1966, pp. 294-299; Ebel, 1951) and estimation of synchronization for selection of units of analysis. The model statements used with aggregated data can be complex, involving many terms that may vary from design to design. For this reason an analysis of variance algorithim[1] is given below, for balance designs, that is more parsimonious than that provided by many commonly used texts (e.g., Winer, 1971, pp. 371-375), to assist the reader with subsequent material and to clarify terminology and notation that is not completely standard.

### Model Statement

Main effect terms are identified by a single alpha character in caps. Nested relationships, if any, are identified by additional alpha characters in brackets next to the term in question, showing what this term is nested within. Interactions are denoted by two or more alpha characters identifying the interesting main effects. The full rank model includes interactions between all combinations of terms, excluding, however, interactions between any terms that share a common alpha character. Terms are ordered by examining the alpha characters denoting terms. If the alpha characters of one term are a subset of the characters of another, the term that is a subset must be placed ahead of the other. Nonnested main effect terms with a greater number of other terms nested within them are listed ahead of the nonnested main effects with fewer other terms nested within them.

### Expected Mean Squares

Expected mean squares (EMS) identify how mean squares are divided into the various components that contribute to the makeup of the mean square. Since expected mean squares are essential for deriving reliability formulas, the following algorithm can be used to derive expected mean squares in the balanced

[1]This algorithm, in similar form but with different notation, should be attributed, to the author's knowledge, to Dr. Melvin Carter, Department of Statistics, Brigham Young University.

case. To see whether the variance components for other terms occur in the expected mean squares for the term in question, the alpha characters of the term in question are examined in relation to the alpha characters of the other terms. If the term in question is a subset of another term, then the complement of the characters is taken. If all of the nonbracketed characters belonging to this complement designate random factors, then the variance component for this other term does occur in the expected mean squares. The coefficient for this variance component, that occurs in the expected mean squares, is found by finding the alpha characters not listed as part of the term. The product of the levels of the main effect terms not listed in this way equals the coefficient.

### Sums of Squares

The sums of squares for any balanced complete-block design, can be readily obtained by: (a) taking the sum over levels of main effects not listed, for the term in question; (b) next squaring and then summing over levels of main effects that are listed; and finally, (c) this sum is then divided by the product of levels of main effects not listed. Then the sum of squares for the term in question is obtained by subtracting all sums of squares of terms that are subsets of the term in question. This includes the $\mu$ term.

### Degrees of Freedom

Degrees of freedom for each term are obtained by taking the product of the levels of the main effects that are listed for the term in question, and then subtracting the degrees of freedom of all terms that are subsets of the term in question. Again this includes the $\mu$ term.

## Data Structure

### Overview

Reliability estimation is dependent upon specifying the structure of the data, which can be identified with an analysis of variance model statement. The following analysis of variance model statement illustrates the type of structure frequently encountered with survey data taken from intact organizational groups. The model statement is used to describe U.S. Army organization, but could equally fit most organizations, and is used as an example throughout the paper.

$$Y = \mu + \underline{A} + \underline{B}(\underline{A}) + \underline{C}(\underline{AB}) + \underline{R} + \underline{AR} + \underline{BR}(\underline{A}) + \underline{CR}(\underline{AB}) + \underline{S}(\underline{ABCR}) + \underline{Q} + \underline{AQ} + \underline{BQ}(\underline{A}) + \underline{CQ}(\underline{AB}) + \underline{RQ} + \underline{ARQ} + \underline{BRQ}(\underline{A}) + \underline{CRQ}(\underline{AB}) + \underline{SQ}(\underline{ABCR}) + \underline{E}(\underline{ABCRSQ}) \quad (1)$$

where, $\underline{A}$ = 1, $\underline{a}$; brigade, random

$\underline{B}$ = 1, $\underline{b}$; battalion, fixed

$\underline{C}$ = 1, $\underline{c}$; company, fixed (except where explicitly specified as random)

$\underline{R}$ = 1, $\underline{r}$; race, fixed

$\underline{S}$ = 1, $\underline{s}$; subjects, fixed or random

$\underline{Q}$ = 1, $\underline{q}$; questionnaire items, fixed or random

$\underline{E}$ = 1, 1; error, random

An Army company consists of approximately 150 soldiers who work together. There are five companies within a battalion and three battalions within a brigade. The hierarchical nature of the organization is specified by the completely-nested hierarchical portion of the design ($\underline{A}$, $\underline{B}$, and $\underline{C}$). Assuming enough units were available, either brigades, battalions or companies could be selected as the unit of analysis. Nesting any number of hierarchical levels is possible. The hierarchical data structure is a very general one that can be applied to most organizations in many societies. It can apply also to generational hierarchies in groups organized along familial lines. Mixed hierarchies can also be examined with families nested within the parental occupational organization(s).

Following the hierarchical part of the design, the term Race ($\underline{R}$) appears, which crosses the hierarchical groups (i.e., it is not nested within them). This crossed term, whether it designates a variable like race (black-white), or rank (supervisor-subordinate), or even generation (parent-child), designates subgroups that represent repeated measurements across the unit of analysis (e.g., companies, families). Repeated measurements across the unit of analysis can be used to examine the interaction between the subgroups that are repeated, by correlating the responses of the subgroup across the units, and when available, across time using cross-lagged panel correlation or path analysis (see Hart, 1978). Interaction between subgroups can be examined over time in this manner. In addition to the single-crossed term Race ($\underline{R}$), other crossed terms designating subgroups with their associated interaction terms are possible, as well as covariates without interactions.

The term representing Questionnaire items ($\underline{Q}$) is crossed with both the nested Subjects term ($\underline{S}$) and the hierarchical terms ($\underline{A}$, $\underline{B}$, $\underline{C}$), which means questionnaire items can be considered repeated measures in two ways--across both subjects and the unit of analysis ($\underline{A}$, $\underline{B}$ or $\underline{C}$). Just one such term is expected, representing survey items. Succeeding terms represent interactions with $\underline{Q}$. Data that is repeated in both ways contain common-method variance (see Campbell & Fiske, 1959) not found in data repeated only across the unit of analysis, so that correlations between variables that are repeated in both ways should be inflated in relation to correlations based on data that is repeated only across the unit of analysis and not across subjects. Data that is repeated in two ways is represented by the ratings of a single subgroup, within the unit of analysis, on two different scales, while data that is repeated in only one way is represented by ratings from two different subgroups on two different scales. Methods of reliability estimation that use the communality between all variables in an analysis (see Kenny, 1975, pp. 897-899; Joreskog & Sorbom, 1979, chap. 4) are not appropriate for data structures, as above, in which correlations are influenced by whether the variable is "repeated" in more than one way. Internal consistency reliabilities are preferable with the above data structure.

Overall, the model can be considered a hierarchical split-plot (or repeated-measures) design. The Q term and interactions with Q represent Within-Subjects variance, while the hierarchical and crossed terms with their interactions represent Behavior-Subjects variance, as found in a split-plot (repeated-measures) design. The between subjects variance can be further divided into two parts--the hierarchical part representing Between-Groups variance, and the crossed term(s) with their interactions representing Within-Groups variance--thus creating the hierarchical split-plot design. Analysis of variance designs like the above generally have more than one error term. For example, the term SQ can be considered an appropriate error term to test within-subjects terms, and S an error to test between-subjects terms. Furthermore, the hierarchical terms C, and B might be considered error terms under some circumstances. Error terms are dictated not only by the model but also by the terms considered fixed and random. The determination of whether a term is fixed or random depends on the sampling plan of the design.

## Sampling Plans

In the previous model statement, Brigades (A) may have been sampled in a random or at least representative fashion, while Battalions (B) and Companies (C) may have been sampled in an exhaustive fashion. Brigades may therefore be random while battalions and companies within brigades are fixed since the population of these units was exhaustively sampled. In the preceding example the nested hierarchical terms B and C were fixed, but in rare cases such terms could be random. For example, if countries were used as a unit of analysis, and in the sampling plan cities were randomly selected to represent countries, with subjects randomly selected within cities, the nested-hierarchical term, cities, could be random as well as subjects.

The Subjects term (S) in the previous example, nested within Companies (C) and Race (R), will be considered fixed or random depending on how exhaustively the population of subjects within companies is sampled. The subjects term is fixed when all soldiers (approximately 150) are sampled, and random when a very small fraction of the company population is sampled. The fixed-random distinction is determined by the sampling fraction ($s/N$, sample size over popuation size), with terms fixed when the ratio is one and random when the ratio is zero. In practice, the subjects terms often will be neither fixed nor random. The company populations are quite small and it's not unusual at all for a sampling plan to call for sampling a fraction of the population (e.g., 1/3) that approaches neither one nor zero. In these cases, the subjects term will be labeled semirandom. The Questionnaire items ($\underline{Q}$) may likewise be considered random if the items in the survey are considered a random selection of a potentially infinite population of items measuring the same concept, or fixed if the items are considered to exhaust the population of interest.

Subjects could be considered random or semirandom and items fixed in a cross-lagged correlation design using groups as the unit of analysis (see Hart, 1978). In this design, a sample of subjects within companies can be selected to represent the whole company population, so subjects are random or semirandom. Cross-lagged correlation looks at time-related changes assuming stationarity--constant item structure over time (Kenny, 1975). In such cases it may often be reasonable to assume items are fixed when looking at time-related changes in this way. Likewise, subjects can be considered fixed and items random in most single-time, survey-feedback designs. In this case, entire company populations are frequently sampled, while items are considered a sample of a larger conceptual population. In this sampling plan subjects become fixed and items random. Of course, in many designs both subjects and items may be random or at least semirandom.

## Reliability Formulas

### Derivation

The sampling plans given above have a direct impact on the appropriate reliability formulas. A requirement for measuring reliability is to divide the variance associated with the unit of analysis into true and error components. The unit of analysis in this case is an aggregated group score instead of an individual response. If the unit of analysis is the Companies term ($\underline{C}$), the expected mean squares for this term show the underlying components that are expected in the make-up of the observed mean square. These underlying components can be divided into true and error variance. This provides a way of allocating the observed company mean square into true and error components. The sampling plan determines which terms are fixed and random. This in turn affects the expected mean squares for the unit of analysis and the allocation of true and error components to the observed mean square, which then affects the reliability formula. Table 1 shows how the expected mean squares in the balanced case change, for selected terms, as a function of whether Subjects ($\underline{S}$) and Questionnaire items ($\underline{Q}$) are considered fixed or random. Reliability is defined as the ratio of true to total variance. The variance component defined as true variance is always that component associated with the unit of analysis--in this case either Companies ($\underline{C}$), Battalions ($\underline{B}$), or Brigades ($\underline{A}$). As indicated by Table 1 there is more than one "error" term when both items and subjects are random. In

Table 1

Balanced Expected Mean Squares with Fixed/Random

Subjects (S) and Items (Q)[1]

| Term | Expected Mean Squares[2] |
|---|---|
| *A* brigade | $bcrsq\sigma^2_{A} + [q\sigma^2_{S}] + (bcrs\sigma^2_{AQ}) + (\sigma^2_{SQ}) + \sigma^2_{E}$ |
| *B*(*A*) battalion | $crsq\sigma^2_{B} + [q\sigma^2_{S}] + (crs\sigma^2_{BQ}) + (\sigma^2_{SQ}) + \sigma^2_{E}$ |
| *C*(*AB*) company | $rsq\sigma^2_{C} + [q\sigma^2_{S}] + (rs\sigma^2_{CQ}) + (\sigma^2_{SQ}) + \sigma^2_{E}$ |
| *S*(*ABCR*) subjects | $q\sigma^2_{S} + (\sigma^2_{SQ}) + \sigma^2_{E}$ |
| *AQ* brigade X items | $bcrs\sigma^2_{AQ} + [\sigma^2_{SQ}] + \sigma^2_{E}$ |
| *BQ*(*A*) battalion X items | $crs\sigma^2_{BQ} + [\sigma^2_{SQ}] + \sigma^2_{E}$ |
| *CQ*(*AB*) company X items | $rs\sigma^2_{CQ} + [\sigma^2_{SQ}] + \sigma^2_{E}$ |
| *SQ*(*ABCR*) subjects X items | $\sigma^2_{SQ} + \sigma^2_{E}$ |

[1] The model and notation are found in the text (see Equation 1). The term *A* is random with *B* and *C* fixed. Subjects (*S*) and Questionnaire Items (*Q*) are either fixed or random. Lower case letters denote the number of levels of the corresponding factors in caps.

[2] When subjects are fixed, terms within brackets are deleted. When questionnaire items are fixed, terms within parentheses are deleted.

general, as the number of main effects following the unit of analysis, that are random, increase, the number of components considered to be error increase dramatically, (see Formula 11, Table 2).

Reliability for the group mean scores is formally defined in Table 2. The expected mean squares, shown in Table 1, for the unit of analysis ($\underline{C}$), are divided by $\underline{rsq}$, the product of the levels that are added to obtain the group means. The divided expected mean squares represent the components expected in the group means, components that vary according to the sampling plan. The component due to the unit of analysis ($\underline{C}$), divided by all components, represents the ratio of true over total variance needed for the reliability definition. Mean square terms are set equal to the corresponding expected mean squares, and then the equations are solved for the variance components. For example, the variance components for definition 3 in Table 2 equal:

$$\sigma^2_{\underline{C}} = (\underline{MS}_{\underline{C}} - \underline{MS}_{\underline{S}}) \ / \ \underline{rsq}; \quad \underline{q}\sigma^2_{\underline{S}} + \sigma^2_{\underline{E}} = \underline{MS}_{\underline{S}}.$$

The mean square estimates of the variance components are substituted for the corresponding variance component in the reliability definition, and then simplified algebraically. This process produced the reliability formulas in Table 2.

The unit of analysis for Formulas (3) through (10) is Companies ($\underline{C}$). When the unit is Battalions ($\underline{B}$) or Brigades ($\underline{A}$), the definitions and reliability formulas are the same as in Table 2, with the following substitutions: (a) $\sigma^2_{\underline{C}}$ becomes $\sigma^2_{\underline{B}}$, or $\sigma^2_{\underline{A}}$; (b) $\sigma^2_{\underline{CQ}}$ becomes $\sigma^2_{\underline{BQ}}$, or $\sigma^2_{\underline{AQ}}$; (c) $\underline{MS}_{\underline{C}}$ becomes $\underline{MS}_{\underline{B}}$, or $\underline{MS}_{\underline{A}}$; and (d) $\underline{MS}_{\underline{CQ}}$ becomes $\underline{MS}_{\underline{BQ}}$, or $\underline{MS}_{\underline{AQ}}$. When the unit of analysis is Battalions ($\underline{B}$), the terms including $\underline{B}$ are substituted, and when the unit is Brigades ($\underline{A}$), $\underline{A}$ is substituted. The error terms in the denominator of the reliability definitions are divided by an additional coefficient $\underline{c}$ for Battalions and $\underline{bc}$ for Brigades.

Estimating reliability involves estimating ratios of variance components. The expectation of these ratios contains a slight positive bias. Winer (1971, pp. 248-249; 282-290) has given a correction for this bias for the standard formulas (Formula 2, Table 2; Formula 26, Table 4). This correction, when extended to any of the formulas in Table 2, has the following form:

$$\frac{\underline{MS}_{unit} - \left(\underline{df}_{error}/(\underline{df}_{error} - 2)\ \underline{MS}_{error}\right)}{\underline{MS}_{unit}} \qquad (12)$$

where, $\underline{MS}_{unit}$ is the mean square for the unit of analysis, $\underline{MS}_{error}$ represents the mean square term(s) measuring error. The term(s) subtracted from $\underline{MS}_{\underline{C}}$ in the numerator of the formulas in Table 2 are error. In words, the correction involves multiplying $\underline{MS}_{error}$ by a correction term that approaches one as the degrees of freedom for error increase. When $\underline{MS}_{error}$ involves more than one mean

Table 2

Reliability Formulas for Mean Scores as a Function of Unit of Analysis and Sampling Plan

| Unit of Analysis | Sampling Plan | Reliability Definition | Formula | Number |
|---|---|---|---|---|
| Subjects (S) | Items random<br>Subjects random | $\frac{\sigma^2_S}{\sigma^2_S + (\sigma^2_{SQ} + \sigma^2_E)/q}$ | $\frac{MS_S - MS_{SQ}}{MS_S}$ | (2) |
| Companies (C) | Items fixed<br>Subjects random | $\frac{\sigma^2_C}{\sigma^2_C + (q\sigma^2_S + \sigma^2_E)/rsq}$ | $\frac{MS_C - MS_S}{MS_C}$ | (3) |
| Companies (C) | Items fixed<br>Subjects semirandom | $\frac{\sigma^2_C + s/N_s\ \sigma^2_S/rs}{\sigma^2_C + (q\sigma^2_S + \sigma^2_E)/rsq}$ | $\frac{MS_C - ((N_s - s)/N_s\ (MS_S - MS_{SQ}) + MS_{SQ})}{MS_C}$ | (4) |
| Companies (C) | Items random<br>Subjects fixed | $\frac{\sigma^2_C}{\sigma^2_C + (rs\sigma^2_{CQ} + \sigma^2_E)/rsq}$ | $\frac{MS_C - MS_{CQ}}{MS_C}$ | (5) |
| Companies (C) | Items semirandom<br>Subjects fixed | $\frac{\sigma^2_C + q/N_q\ \sigma^2_{CQ}/q}{\sigma^2_C + (rs\sigma^2_{CQ} + \sigma^2_E)/rsq}$ | $\frac{MS_C - ((N_q - q)/N_q\ (MS_{CQ} - MS_{SQ}) + MS_{SQ})}{MS_C}$ | (6) |
| Companies (C) | Items random<br>Subjects random | $\frac{\sigma^2_C}{\sigma^2_C + (q\sigma^2_S + rs\sigma^2_{CQ} + \sigma^2_{SQ} + \sigma^2_E)/rsq}$ | $\frac{MS_C - MS_S - MS_{CQ} + MS_{SQ}}{MS_C}$ | (7) |
| Companies (C) | Items random<br>Subjects semirandom | $\frac{\sigma^2_C + s/N_s\ \sigma^2_S/rs}{\sigma^2_C + (q\sigma^2_S + rs\sigma^2_{CQ} + \sigma^2_{SQ} + \sigma^2_E)/rsq}$ | $\frac{MS_C - ((N_s - s)/N_s\ (MS_S - MS_{SQ}) + MS_{CQ})}{MS_C}$ | (8) |

Table 2 (continued)

| Unit of Analysis | Sampling Plan | Reliability Definition | Formula | Number |
| --- | --- | --- | --- | --- |
| Companies (C) | Items semirandom<br>Subjects random | $\frac{\sigma^2_C + q/N_q\ \sigma^2_{CQ}/q}{\sigma^2_C + (q\sigma^2_S + rs\sigma^2_{CQ} + \sigma^2_{SQ} + \sigma^2_E)/rsq}$ | $\frac{MC_C - ((N_q - q)/N_q\ (MS_{CQ} - MS_{SQ}) + MS_S)}{MS_C}$ | (9) |
| Companies (C) | Items semirandom<br>Subjects semirandom | $\frac{\sigma^2_C + s/N_s\ \sigma^2_S/rs + q/N_q\ \sigma^2_{CQ}/q}{\sigma^2_C + (q\sigma^2_S + rs\sigma^2_{CQ} + \sigma^2_{SQ} + \sigma^2_E)/rsq}$ | $\frac{MS_C - ((N_s - s)/N_s\ (MS_S - MS_{SQ}) + (N_q - q)/N_q\ (MS_{CQ} - MS_{SQ}) + MS_{SQ})}{MS_C}$ | (10) |
| Battalion (B) | Items random<br>Subjects random | $\frac{\sigma^2_B}{\sigma^2_B + (rsq\sigma^2_C + q\sigma^2_S + crs\sigma^2_{BQ} + rs\sigma^2_{CQ} + \sigma^2_{SQ} + \sigma^2_E)/crsq}$ | $\frac{MS_B - MS_C - MS_{BQ} + MS_{CQ}}{MS_B}$ | (11) |

Note. All formulas are based on the analysis of variance model statement given in Equation 1, except for Formula (2) which is based on the following model: $Y = S + Q + SQ + E(SQ)$, with $S$ and $Q$ random. The terms $S$ and $Q$ are defined as in Equation 1. The terms $N_s$ and $N_q$ refer to the population size for (a) the number of subjects in each company, and (b) the number of items in the population of interest, respectively. Brigade is considered random, and battalion and company fixed in all cases except Formula (11) where company is considered random as well, to show how the formulas change. Formulas (4) and (6) assume $\sigma^2_{SQ} = 0$, so that $MS_{SQ} = \sigma^2_E$.

square term, the adjusted degrees of freedom for these several terms are found by referring to Formula (24) given later. For all practical purposes the positive bias in the reliability formulas in Table 2 is negligible with as many degrees of freedom for $\underline{MS}_{error}$ as is customary with organizational surveys.

Another bias may be more serious. As with any analysis of variance design, if significant terms are omitted from the model statement, these omitted terms will artificially inflate $\underline{MS}_{error}$. Reliability will be underestimated to the extent significant terms are omitted from the model statement. For example, omitting Race ($\underline{R}$) when it, or its interactions, are significant, increases the size of $\underline{MS}_{\underline{S}}$. It is desirable to specify model statements that capture the structure of the data as completely as possible even if this creates model statements with large numbers of terms.

## Interpretation

The reliabilities are internal consistency measures of reliability. As such they represent reliability at any one discrete point in time. At this point in time the reliabilities measure the extent to which the researcher would expect to obtain the same thing if the measurement process were repeated. They estimate the correlation between the mean scores, for the unit of analysis, and another set of mean scores that would be expected if the measurement process had been repeated at the same time. The reliability would also be considered an estimate of the correlation between the observed sample means and the means that would have been obtained if the entire population of subjects/items had been measured.

The sampling plans differ for different reliability formulas. Sampling is conducted without replacement (i.e., no respondent takes the survey twice at one time) which creates the practical effect of sampling from a population that can be considered finite. When subjects are fixed, the "observations" that make up the variation due to subjects $\sigma^2_{\underline{S}}$, remain the same in the hypothetical new sample as they were in the observed sample, and when subjects are semirandom the proportion of these elements in each group that remain the same equals $\underline{s}$ / $\underline{N}_{\underline{s}}$ (sample over population size). Likewise, when items are fixed, the "observations" due to the component $\sigma^2_{\underline{CQ}}$ are identical in the observed and hypothetical new sample, and in the semirandom case the proportion of elements that are the same equals $\underline{q}$ / $\underline{N}_{\underline{q}}$. When the sample size equals the population size (i.e., the term is fixed), the same scores are selected twice, the mean scores are measured without error, and the reliability is perfect. When a term is semirandom, the hypothetical new sample will contain $\underline{n}$ / $\underline{N}$ elements in common with the old sample and the population. When a term is random, none of the elements that make up that component remains the same in the new sample or population. Declaring a term fixed or random, then, is the same as assuming the elements that go into a particular variance component either change or do not change from the observed sample to a hypothetical new one or to the population. They do not change if the sample size equals the population size.

## Relationship Between Formulas

In fact, there is a close connection between average intercorrelation, and reliability as computed by Cronbach's alpha, and analysis of variance. Cronbach's alpha is identical to the Spearman-Brown prediction formula applied to the average intercorrelation between items (see Ebel, 1951). Formula 1 in Table 2 differs from Cronbach's alpha only in that analysis of variance, with its attendant assumptions, is used to estimate the average intercorrelation between items (see Formula 26, Table 4). This estimate of the average intercorrelation (Formula 26), when corrected by the Spearman-Brown prediction formula, equals Formula 2.

When computing reliability for aggregated scales researchers typically compute Cronbach's alpha on group means, computed separately for each item, which is the same as computing the average intercorrelation between these item means, and adjusting the average correlation with the Spearman-Brown prediction formula. This is closely approximated by Formula 5, Table 2. The average intercorrelation between company mean scores for each item is estimated by Formula 27, Table 4. When this analysis of variance estimate of the average intercorrelation is corrected by the Spearman-Brown prediction formula it equals Formula 5. The use of Cronbach's alpha to estimate the reliability of group mean scores requires the same sampling assumptions as does Formula 5--subjects fixed and items random. When subjects are sampled from large intact organizational groups, Formula 5 is not appropriate and neither is Cronbach's alpha. For example, Taylor and Bowers (1972) used Cronbach's alpha both on exhaustive and ten percent samples of subjects. Formula 5 should have given way to Formula 8 with the ten percent sample if the assumption of random items had been made.

A comparison of Formulas (2) and (3), Table 2, shows an interesting relationship between variance components. When individuals are used as the unit of analysis, the between subjects variance $\sigma^2_{\underline{S}}$ represents true variance, but when companies are the unit, and subjects are random, as in Formula 3, the terms $\sigma^2_{\underline{S}}$

represents error variance. It is true that the subjects components $\sigma^2_S$ are not identical in the two cases since the models differ, but they are very similar. The subjects mean square ($MS_S$) in Formula 3 has been reduced compared to the subjects mean square $MS_S$ in Formula 2, to the extent that other "between subjects" terms from the model in Equation 1 are significant, but otherwise the terms are the same. Maximizing the variance between subjects will increase reliability as measured by Formula 2, but can decrease it as measured by Formula 3. In constructing the Survey of Organizations (see Taylor & Bowers, 1972), "between subjects" variance was maximized by such techniques as (a) positive wording of all questions, (b) contiguous placement of items from the same scale, (c) positive response alternatives lined up on the same side of the scale, and (d) selection of items with large "between subjects" distributions. These techniques will maximize reliability as measured by Formula 2. The previous techniques seem to maximize subject differences by increasing variance due to response sets. If this is the case, this subject variance would be expected to inflate $MS_S$ as error in Formula 3. It is possible that these techniques also reduce $\sigma^2_E$ so it may not always increase $MS_S$ as error. In Formula 3 we wish to maximize $MS_C$ in relation to $MS_S$. The preceding technique used in Survey of Organizations could easily, but not necessarily, increase $MS_S$ in relation to $MS_C$, reducing reliability. Since the Survey of Organizations and others like it, use intact organizational groups as units, Formula 3 rather than 2 is most appropriate and should be used when subjects alone are random.

Formulas 2 and 5 have generally been used to establish reliability for organizational surveys. It should be apparent from Table 2 that there is no necessary relationship between reliability as measured by Formula 5 and 3. Furthermore, there may sometimes be a negative relationship between reliability as measured by Formula 2 and 3. Organizational Surveys that claim to have well established reliabilities, using Formulas 2 or 5, have not established reliability at all for the situations in which Formulas 3, 4, 7, 8, 9 or 10 are most appropriate. In fact, it is reasonable to suppose that many of these "well established reliabilities" will not prove to be reliable at all as measured by Formula 3, since no attempt has been made, using pretest samples to select items that discriminate well between group units, while a corresponding effort has been made to find items that have high intercorrelations. It is important to find which scales are in fact reliable using appropriate formulas. Research in this direction may require a reassessment of the reliabilities of the scales used in organizational research, as well as interpretations of results in this area.

## Reliability for Record Data

Frequently variables representing group units of analysis are not measured by survey but can be found in the form of frequency counts of events within the group, that occurred during a given time period. Often these frequency counts are expressed in the form of rates (e.g., per 1000) or percentages. The use of rates or percentages is generally not a good idea when the variables are to be correlated, since this creates the attendant problems of index correlation (see McNemar, pp. 180-182). A better approach is to use the raw frequency counts, and partial out the effects of sample size (Cronbach & Furby, 1970). Reliability for such frequency counts can be computed using analysis of variance, with the group size variable used as a covariate. The model in this case differs slightly from that shown in Equation (1). The following model defines the structure of the data in the case with three levels of hierarchy:

$$Y = \underline{A} + \underline{B}(\underline{A}) + \underline{C}(\underline{AB}) + \underline{D} + \underline{AD} + \underline{BD}(\underline{A}) + \underline{CD}(\underline{AB}) + \underline{E}(\underline{ABCD}) \quad (13)$$

where, $\underline{A} = 1, \underline{a}$; brigade, random

$\underline{B} = 1, \underline{b}$; battalion, fixed

$\underline{C} = 1, \underline{c}$; company, fixed

$\underline{D} = 1, \underline{d}$; generally dichotomous split of frequencies, random

$\underline{E} = 1, 1$; error, random

The addition of another crossed term like Race ($\underline{R}$), that is fixed, does not affect the reliability definition or formula, so it was omitted. In addition to the above model the group size variable can be added as a covariate. The term D can represent either a random dichotomous split, or a dichotomous split that controls for a variable like time (e.g., one level represents events that occurred on odd numbered days and the other level events that occurred on even numbered days for the time period in question). The split may have to be random when the time variable is not available on a case by case basis. The fact that a random split is possible means that an internal consistency reliability can be computed when only frequency counts are available for each group. Researchers often assume it is not possible to compute reliability in this case. The reliability definition and formula are given as follows:

$$\frac{\sigma^2_{\underline{C}}}{\sigma^2_{\underline{C}} + (\sigma^2_{\underline{CD}} + \sigma^2_{\underline{E}}) / \underline{d}} = \frac{\underline{MS}_{\underline{C}} - \underline{MS}_{\underline{CD}}}{\underline{MS}_{\underline{C}}} \quad (14)$$

When random splits within groups are necessary to obtain the observations for the term D, greater stability in the reliability estimates can be obtained by a jacknife procedure in which $\underline{MS}_{\underline{CD}}$ in Formula (14) is estimated several times using different random splits each time. The different estimates can then be averaged prior to using the averaged estimate in Formula (14). When the term $\underline{D}$ is fixed the record variable in question is considered to be measured without error and an estimate of reliability is not needed. This would occur if (a) the researcher was willing to limit generalizations to that particular variable alone, and (b) the frequencies of that variable were a census rather than sample of the relevant events.

## Significance Tests

### Difference of Reliability from Zero

It is important to ask if it is possible to detect a significant amount of true variance at all, i.e., is the reliability coefficient significantly different from zero. One form in which this test can be made is to compare total to error variance, forming an *F* ratio, to see if a detectable amount of true variance exists. The form of the *F* test differs slightly from the reliability ratio (true over total variance), but provides a test with the same components. The Test definitions and *F* tests for reliability Formulas 3 through 10 are shown in Table 3. The error terms in the denominators of the *F* ratios in Table 3 can be found in different form as the quantity subtracted from $MS_C$ in the numerator of the reliability formulas in Table 2. The error terms are expressed in different form in Table 3 because tests (17) through (23) are quasi-*F* tests, i.e., tests involving more than two mean square terms in the *F* test. In this case, the *F* test is an approximation which is obtained by adjusting the degrees of freedom for both the numerator and denominator separately, by the formula given in Satterthwaite (1946):

$$\underline{df}\text{ adj.} = \frac{\left(\underline{a}_1(\underline{MS}_1) + \underline{a}_2\ (\underline{MS}_2) + \ldots\right)^2}{\dfrac{\left(\underline{a}_1(\underline{MS}_1)\right)^2}{\underline{df}_1} + \dfrac{\left(\underline{a}_2(\underline{MS}_2)\right)^2 \ldots}{\underline{df}_2}} \qquad (24)$$

where, $\underline{MS}_1$ and $\underline{MS}_2$ are independent mean squares, and $\underline{a}_1$ and $\underline{a}_2$ are the coefficients for the mean squares. The mean squares in Table 3 are shown in a form that gives separate coefficients for each mean square as required by Formula 24. In the case where group size is unbalanced, and the coefficients, $\underline{a}_i$, vary from company to company, the quantity $\underline{a}_i\ \underline{MS}_i$ can be obtained most accurately by weighting individual scores as appropriate (e.g., Formula 42, as described later).

### Difference Between Reliabilities

In some situations it is important to know whether reliabilities are significantly different from each other. For example, using cross-lagged panel correlation (Kenny, 1975), it is important to know whether reliability changes over time. When reliability changes, corrections for reliability shifts are made. A statistical test for reliability shifts is desirable and can be made when the reliabilities are expressed in the form of F ratios as shown previously in Table 3, and the assumption is made that the mean square terms are independent. In the case where measurements are made on group units at more than one point in time, with different subjects sampled on each occasion, the samples involve the same group populations but different subjects. In analysis of variance terms, the measurements are repeated across companies, but not across subjects. The mean square terms under these conditions approximate independence. The bias due to lack of independence is loss of power. Degrees of freedom are large enough so that power is not low in any case. Following Winer (1971, pp. 245-247), hypotheses related to the equality of two F ratios can be tested as follows:

$$F_L > (F_S)\,[F_\alpha\,(df\text{ numerator, }df\text{ denominator})] \qquad (25)$$

where, $F_L$ and $F_S$ represent reliabilities in the form of F ratios as shown in Table 3; $F_L$ representing the larger F ratio and $F_S$ the smaller. To obtain $F_\alpha$, the degrees of freedom in the numerator and denominator should correspond to degrees of freedom in the numerator and denominator of $F_L$ and $F_S$. The degrees of freedom for $F_L$ should approximately equal those for $F_S$ for the test to be valid. When quasi-F ratios are used, the degrees of freedom for $F_\alpha$ should correspond to adjusted degrees of freedom as given in Equation (24). The test should be used with some caution with quasi-F ratios.

## Sample Size Requirements

Organizational research is costly and time consuming. For these reasons, it is important to be able to estimate ahead of time the sample sizes needed to obtain specified levels of reliability desired by the researcher. How many subjects within each group, and how many items in a scale are needed to obtain a specified level of reliability, say .75, as measured by the formulas in Table 2? Estimates of the mean square terms in Table 2 can be obtained from a pretest sample, and from the pretest sample the number of subjects and items that are needed for a specified level of reliability can be estimated.

The way this problem has been solved in the standard case where individuals are the unit of analysis, has been to estimate the reliability of a single score (Formula 26, Table 4) which is related to the reliability of the average score (Formula 2, Table 2) in terms of the Spearman-Brown prediction formula. Solving the Spearman-Brown prediction formula for the sample size, tells how many items must be added to obtain the desired reliability (see Winer 1971, p. 287). This same approach was used in Table 4 for other formulas. However, when the unit of analysis involves a group, the reliability of single scores involves contingencies: the reliability of a single item given the same number of subjects as was

Table 3

Statistical Significance of Reliability Coefficients[a]

| Reliability Formula[b] | Test Definition | $\underline{F}$ Test[c] | Formula Number |
|---|---|---|---|
| 2 | $\dfrac{q\sigma^2_S + \sigma^2_{SQ} + \sigma^2_E}{\sigma^2_{SQ} + \sigma^2_E}$ | $\dfrac{MS_S}{MS_{SQ}}$ | (15) |
| 3 | $\dfrac{rsq\sigma^2_C + q\sigma^2_S + \sigma^2_E}{q\sigma^2_S + \sigma^2_E}$ | $\dfrac{MS_C}{MS_S}$ | (16) |
| 4 | $\dfrac{rsq\sigma^2_C + q\sigma^2_S + \sigma^2_E}{((N_s - s)/N_s\, q\sigma^2_S) + \sigma^2_E}$ | $\dfrac{MS_C}{(N_s - s)/N_s\, MS_S + s/N_s\, MS_{SQ}}$ | (17) |
| 5 | $\dfrac{rsq\sigma^2_C + rs\sigma^2_{CQ} + \sigma^2_E}{rs\sigma^2_{CQ} + \sigma^2_E}$ | $\dfrac{MS_C}{MS_{CQ}}$ | (18) |
| 6 | $\dfrac{rsq\sigma^2_C + rs\sigma^2_{CQ} + \sigma^2_E}{((N_q - q)/N_q\, rs\sigma^2_{CQ}) + \sigma^2_E}$ | $\dfrac{MS_C}{(N_q - q)/N_q\, MS_{CQ} + q/N_q\, MS_{SQ}}$ | (19) |
| 7 | $\dfrac{rsq\sigma^2_C + q\sigma^2_S + rs\sigma^2_{CQ} + \sigma^2_{SQ} + \sigma^2_E}{q\sigma^2_S + rs\sigma^2_{CQ} + \sigma^2_{SQ} + \sigma^2_E}$ | $\dfrac{MS_C}{MS_S + MS_{CQ} - MS_{SQ}}$ | (20) |

Table 3 (continued)

| Reliability Formula[b] | Test Definition | $\underline{F}$ Test[c] | Formula Number |
|---|---|---|---|
| 8 | $\dfrac{rsq\sigma^2_C + q\sigma^2_S + rs\sigma^2_{CQ} + \sigma^2_{SQ} + \sigma^2_E}{\left((N_s - s)/N_s\, q\sigma^2_S\right) + rs\sigma^2_{CQ} + \sigma^2_{SQ} + \sigma^2_E}$ | $\dfrac{MS_C}{(N_s - s)/N_s\, MS_S - (N_s - s)/N_s\, MS_{SQ} + MS_{CQ}}$ | (21) |
| 9 | $\dfrac{rsq\sigma^2_C + rs\sigma^2_{CQ} + q\sigma^2_S + \sigma^2_{SQ} + \sigma^2_E}{\left((N_q - q)/N_q\, rs\sigma^2_{CQ})\right) + q\sigma^2_S + \sigma^2_{SQ} + \sigma^2_E}$ | $\dfrac{MS_C}{(N_q - q)/N_q\, MS_{CQ} - (N_q - q)/N_q\, MS_{SQ} + MS_S}$ | (22) |
| 10 | $\dfrac{rsc\sigma^2_C + rs\sigma^2_{CQ} + q\sigma^2_S + \sigma^2_{SQ} + \sigma^2_E}{\left((N_s - s)/N_s\, q\sigma^2_S\right) + \left((N_q - q)/N_q\, rs\sigma^2_{CQ}\right) + \sigma^2_{SQ} + \sigma^2_E}$ | $\dfrac{MS_C}{(N_s - s)/N_s\, MS_S + (N_q - q)/N_q\, MS_{CQ} - (1 - s/N_s - q/N_q)\, MS_{SQ}}$ | (23) |

[a]The test is for the significance of the reliability coefficient from zero. It is defined in terms of true plus error (total) variance over error variance alone. It will answer the question of whether it is possible to detect any true variance at all. The component $\sigma^2_{SQ}$ is assumed zero, $MS_{SQ} = \sigma^2_E$, for Formulas (17) and (19).

[b]The numbers refer to the reliability formulas in Table 2.

[c]When two or more mean squares are found in the denominator, the $\underline{F}$ test is an approximation which is obtained by adjusting the degrees of freedom for the denominator by Formula 24.

Table 4

Reliability Formulas for Single Scores as a Function of Unit Of Analyses and Sampling Plan

| Unit of Analysis | Sampling Plan | Score Estimated | Reliability Definition | Formula | Number |
|---|---|---|---|---|---|
| Subjects (S) | Items random<br>Subjects random | Single item | $\frac{\sigma^2_S}{\sigma^2_S + \sigma^2_{SQ} + \sigma^2_E}$ | $\frac{MS_S - MS_{SQ}}{MS_S + (q - 1)\, MS_{SQ}}$ | (26) |
| Companies (C) | Items random<br>Subjects fixed | Single item/ subjects | $\frac{\sigma^2_C}{\sigma^2_C + (rs\sigma^2_{CQ} + \sigma^2_E)/rs}$ | $\frac{MS_C - MS_{CQ}}{MS_C + (q - 1)\, MS_{CQ}}$ | (27) |
| Companies (C) | Items fixed<br>Subjects random | Single subject/ items | $\frac{\sigma^2_C}{\sigma^2_C + (q\sigma^2_S + \sigma^2_E)/rq}$ | $\frac{MS_C - MS_S}{MS_C + (s - 1)\, MS_S}$ | (28) |
| Companies (C) | Items random<br>Subjects random | Single item/ subjects | $\frac{\sigma^2_C}{\sigma^2_C + (q\sigma^2_S + rs\sigma^2_{CQ} + \sigma^2_{SQ} + \sigma^2_E)/rs}$ | $\frac{MS_C - MS_S - MS_{CQ} + MS_{SQ}}{MS_C + (q - 1)\, MS_S + (q - 1)\, MS_{CQ} - (q - 1)\, MS_{SQ}}$ | (29) |
| Companies (C) | Items random<br>Subjects random | Single subject/ items | $\frac{\sigma^2_C}{\sigma^2_C + (q\sigma^2_S + rs\sigma^2_{CQ} + \sigma^2_{SQ} + \sigma^2_E)/rq}$ | $\frac{MS_C - MS_S - MS_{CQ} + MS_{SQ}}{MS_C + (s - 1)\, MS_S + (s - 1)\, MS_{CQ} - (s - 1)\, MS_{SQ}}$ | (30) |

*Note*. All formulas are related to the corresponding formulas in Table 2 in terms of the Spearman-Brown prediction formula, which takes the following form for sample size:

$$n = \frac{R_W\,(1 - R_1)}{R_1\,(1 - R_W)}$$

where, $R_W$ equals the reliability the researcher wants, $R_1$ equals the reliability of a single score as given in this table and $n$ equals sample size required. If $q_2$ equals the number of items required, $s_2$ equals the number of subjects required in each group, $q_1$ equals the number of items in the pretest, and $s_1$ equals the number of subjects within groups in the pretest, $n = q_2$, given $s_1 = s_2$, or $n = s_2$, given $q_1 = q_2$.

found in the pretest sample, or the reliability of a single subject given the same number of items as found in the pretest questionnaire. Given these contingencies, the formulas in Table 4 are related to the corresponding formulas in Table 2, in terms of the Spearman-Brown prediction formula. The corresponding formulas are those with the same unit of analysis and sampling plan. As shown in Table 4, sample size can then be found from the Spearman-Brown formula. Formula (28) and the Spearman-Brown formula can be expressed in more convenient form by solving (28) in terms of the $F$ ratio, $F = MS_C/MS_S$, and substituting this into the Spearman-Brown formula. The number of subjects needed in each group ($s_2$) can then be found as follows:

$$s_2 = \frac{R_W s_1}{F\ (1 - R_W) + R_W - 1}, \qquad (31)$$

where, $R_W$ equals the reliability desired, $s_1$ the sample size in each pretest group and $F = MS_C\ /\ MS_S$.

The problem with using formulas (27) through (31) to estimate sample size requirements is that the number of subjects needed ($s_2$) can only be estimated, given that the number of items to be used in the final questionnaire ($q_2$) equals the number of items ($q_1$) in the pretest sample. The number of items needed in the questionnaire ($q_2$) can only be estimated, given that the number of subjects to be used in the final sample ($s_2$) equals the number used in the pretest ($s_1$). Also, if the unit of analysis is at a higher level than companies, the pretest sample must be assumed to have the same subordinate group structure as in the final sample. Another serious problem is that the preceding approach does not work for some formulas--when subjects or items are semirandom. There are problems with the concept of a single-score reliability in the semirandom case.

The sample size requirement problem was solved for all formulas without any contingencies, by estimating variance components from pretest data independently of the number of subjects or items in the pretest, substituting the sample sizes desired, $s_2$, $q_2$, for pretest coefficients $s_1$ and $q_1$, where they appeared in the reliability definitions, and then solving for $s_2$ and $q_2$. The required formulas are shown in Table 5. From Table 5, the number of subjects or items required for any formula in Table 2 can be estimated from pretest data without any contingencies. For example, a researcher can estimate the number of subjects required ($s_2$), given that $X$ number of items are added to a scale over what existed in the pretest. Similarly, the number of items ($q_2$) can be estimated, given that the sample size within each group in the final sample is larger than it was in the pretest. Of course, the assumption is made that the items that are added are intercorrelated together to the same degree as pretest items above, and subjects

Table 5

Formulas for Determining Sample-Size Requirements

from Pretest Data

| Reliability Formula Defining $R_w$[a] | Sample Size Formulas: Number of Subjects | Sample Size Formulas: Number of Items | Formula Number |
|---|---|---|---|
| 3 | $s_2 = A/C$ | ----- | (32) |
| 4 | $s_2 = A/(C + H)$ | ----- | (33) |
| 5 | ----- | $q_2 = B/D$ | (34) |
| 6 | ----- | $q_2 = B/(D + I)$ | (35) |
| 7 | $s_2 + A/(E - G)$ | $q_2 = B/(E - F)$ | (36) |
| 8 | $s_2 = A/(E - G + H)$ | $q_2 = B/(E - F + H)$ | (37) |
| 9 | $s_2 = A/(E - G + I)$ | $q_2 = B/(E - F + I)$ | (38) |
| 10 | $s_2 = A/(E - G + H + I)$ | $q_2 = B/(E - F + H + I)$ | (39) |

*Note.* $A = s_1 R_w \left(MS_S - (q_2 - q_1)/q_2\ MS_{SQ}\right)$

$B = q_1 R_w \left(MS_{CQ} - (s_2 - s_1)/s_2\ MS_{SQ}\right)$

$C = MS_C\ (1 - R_w) - MS_S (1 - R_w)$

$D = MS_C\ (1 - R_w) - MS_{CQ}\ (1 - R_w)$

$E = MS_C\ (1 - R) - MS_S\ (1 - R_w) - MS_{CQ}\ (1 - R_w) + MS_{SQ}\ (1 - R_w)$

$F = R_w s_1/s_2\ (MS_S - MS_{SQ})$

$G = R_w q_1/q_2\ (MS_{CQ} - MS_{SQ})$

$H = s_1/N_s\ (MS_S - MS_{SQ})$

$I = q_1/N_q\ (MS_{CQ} - MS_{SQ})$

Table 5 (continued)

$R_w$ is the value of the reliability that the researcher wants to obtain in a new sample. The symbol $s_2$ refers to the number of subjects within each group that is needed to obtain the desired reliability $R_w$, while $s_1$ is the pretest sample size within each group. Similarly, $q_2$ refers to the number of items needed to obtain the stated $R_w$, while $q_1$ is the number of items in the pretest. $N_s$ is the population size within each company, while $N_q$ is the size of the population of items. The mean square terms are based on the pretest data using the original model given in Formula (1). The assumption that $\sigma^2_{SQ} = 0$ must be made for Formulas (32), (33), (34), and (35). When $A$ or $B$ is the unit of analysis $MS_A$ or $MS_B$ is substituted for $MS_C$, and $MS_{AQ}$ or $MS_{BQ}$ for $MS_{CQ}$.

[a]The numbers refer to the reliability formulas found in Table 2.

added discriminate between groups to the same extent as in the pretest. The Formulas in Table 5 can be used for any of the units of analysis A, B or C, without contingencies, using the appropriate substitutions given in this table.

Adding items to a survey scale will increase reliability as defined by Formulas (3) and (4), only to a limited extent (i.e., increasing the coefficients of $\sigma^2_C$ and $\sigma^2_S$ in relation to $\sigma^2_E$), and likewise increasing the number of subjects will increase Formulas (5) and (6) only to a limited extent (i.e., increasing the coefficients of $\sigma^2_C$ and $\sigma^2_{CQ}$ in relation to $\alpha^2_E$). Therefore, it is not meaningful to solve the equations for items ($q_2$) for Formulas (32) and (33), or for subjects ($s_2$) for formulas (34) and (35). Negative estimates from any of the formulas in Table 5 mean an infinity of subjects or items would be needed to obtain the requisite reliability, i.e., the desired level of reliability can't be obtained by adding to the sample size.

## Unbalanced Designs

### Effects on Formulas

The derivation of all the previous formulas has been based on the assumption of a balanced design, i.e., equal sample and group sizes across levels of all factors. This, of course, rarely occurs in intact organizations that are of interest here. The impact of unbalanced designs on the expected mean squares, for the model at Equation (1), is shown in Table 6. When balanced formulas are used to calculate the mean squares for the model at Equation 1 when the model is not balanced, the resulting mean squares contain elements of variance components from a variety of extra terms. A comparison of Table 6 and 1 shows additional components or elements of these components, added by unbalance. How the confounding is handled depends entirely on the hypotheses being tested. For purposes of reliability estimation, researchers do not wish to generalize to hypothetical organizations in which groups are all the same size, with equal numbers of, say, blacks and whites in each. Such a balanced hypothesis is clearly irrelevant and inappropriate for intact organizations. Generalizations are made to the intact organization where subgroups vary. In the intact organization the crossed term Race (R) and the subordinate hierarchical terms B(A), and C(AB) are fixed. When these terms are all fixed, it is appropriate to consider all confounded elements added by imbalance to the "between people" components of $MS_A$, $MS_B$, or $MS_C$ as true variance, since that sort of confounding exists naturally in the intact organization to which generalizations are being made. However, when questionnaire items (Q) are considered random, all confounded elements added by unbalance to the "within people" components of $MS_A$, $MS_B$ or $MS_C$ can best be considered error. These confounded elements all represent interactions with the random term Q. Since Q is random, items change from one sample to another, and so would interactions with Q, which suggests these confounded elements should be considered error. When the preceding allocation of

Table 6

Unbalanced Expected Mean Squares

| Model Terms | Between People | | | | | | | | Within People | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | B(A) | C(AB) | R | AR | BR(A) | CR(AB) | S(ABCR) | Q | AQ | BQ(A) | CQ(AB) | RQ | ARQ | BRQ(A) | CRQ(AB) | SQ(ABCR) |
| A | abc | ac | ac | abc | abc | abc | abc | abc | | ab | a | a | ab | ab | ab | ab | abc |
| B(A) | | abc | ac | abc | abc | abc | abc | abc | | | ab | a | ab | ab | ab | ab | abc |
| C(AB) | | | abc | abc | abc | abc | abc | abc | | | | ab | ab | ab | ab | ab | abc |
| S(ABCR) | | | | | | | | abc | | | | | | | | | abc |
| AQ | | | | | | | | | | abc | ac | ac | abc | abc | abc | abc | abc |
| BQ(A) | | | | | | | | | | | abc | ac | abc | abc | abc | abc | abc |
| CQ(AB) | | | | | | | | | | | | abc | abc | abc | abc | abc | abc |
| SQ(ABCR) | | | | | | | | | | | | | | | | | abc |

*Note*. The model is based on Equation 1. The expected mean squares for the terms at left are found in the unbalanced case by looking along the rows for common letters that represent the following conditions: (a) confounding Between Groups, confounding with Race (R), Q random; (b) no confounding Between Groups, confounding with Race (R), Q random; (c) confounding Between Groups, confounding with Race (R). Q fixed. In each case Subjects (S) is considered random.

confounded elements is made between true and error variance, the reliability formulas, tests, and sample size requirements given previously in Tables 2, 3, 4 and 5 remain unchanged. However, it should be recognized that reliability and test definitions contain additional confounded elements as shown in Table 6.

An additional problem remains for hypothesis testing with unbalanced designs. Mean square terms are no longer independent--an assumption required for numerators and denominators of $F$ tests. Tests should be made with caution when unbalance is severe. This problem is not unique to reliability estimation, and is frequently encountered in unbalanced analysis of variance designs.

## Weighting Scores

Unbalanced designs and sampling requirements often necessitate weighting individual scores in order to appropriately estimate reliability. Since sample size affects reliability, as shown previously, weights must be applied in a manner that does not affect the total sample size. Weights are appropriate in the following three situations.

First, using a stratified sampling plan, the crossed term Race ($R$) might not be sampled in proportion to company racial populations. Blacks might be sampled at a higher rate in order to get a sufficient minority sample size. When estimating a total company score, ignoring race, the individual scores within each company need to be weighted to estimate what would have been obtained without disproportionate sampling. In this case the individual scores within each company are weighted according to the following formula:

$$W_{B_i} = \frac{N_{B_i}}{N_{T_i}} \cdot \frac{n_{T_i}}{n_{B_i}}, \tag{40}$$

where, $W_{B_i}$ represents the weight for black subjects in company i, $N_{B_i}$ and $N_{T_i}$ represent, respectively, the black and total population sizes in company i, and $n_{B_i}$ and $n_{T_i}$ represent, respectively, the black and total survey sample sizes. To obtain the weight for white subjects in company i, $N_{W_i}$ and $n_{W_i}$ representing, respectively, the population and sample sizes of whites in company i are substituted to replace $N_{B_i}$ and $n_{B_i}$ in Formula 40.

A second reason for weighting individual scores is to insure that the units of analysis are weighted equally. Since each unit, as a data point, is weighted equally when used in correlation or other statistics, each unit should be weighted equally when estimating reliability. Typically, equal sample sizes are obtained from groups at the level intended for use as the unit of analysis, providing equal weights. However, weights equal at this level will not be equal at another level when hierarchical levels are confounded. Furthermore, a simple random sample may have been used which will produce unequal weights when group sizes differ. In these cases, individual scores within each group or company are weighted as follows:

$$W_i = \frac{N_i}{N_T} \cdot \frac{n_T}{n_i} \tag{41}$$

where $W_i$ is the weight given individual responses within each company, $N_i$ and $n_i$ represent, respectively, the population and sample size for company i, and $N_T$ and $n_T$ represent, respectively, the population and sample totals for all companies combined.

A third reason for weighting individual scores, is to accurately estimate the error terms in Table 2 when subjects are considered semirandom (Formulas 4, 8 and 10). Each unit should be weighted equally in terms of sample size, but the company population sizes are unlikely to be equal also. That means the sampling term $(N_s - s) / N_s$ found in Table 2 will differ from company to company. In order to accurately estimate the error terms $MS_S$ and $MS_{SQ}$ for these semirandom formulas, individual scores within each company should be weighted as follows:

$$W_i = (N_{s_i} - s_i) / N_{s_i}, \tag{42}$$

where, $W_i$ equals the weight in each company and $N_{s_i}$ and $s_i$ represent, respectively, the population and sample sizes in each company. $MS_S$ and $MS_{SQ}$, obtained from scores weighted by (42) are substituted in Formulas 4, 8, and 10 to replace the corresponding terms that are multiplied by $(N_s - s) / N_s$. The other means square terms are estimated without weighting.

The three types of weighting given in Formulas (40), (41), and (42) may be used separately or together in any combination as appropriate. The weights given in (40) and (41) maintain the original sample sizes as required.

## Synchronization Measures

### Making the Measures Comparable

Synchronization measures, are shown in Table 7. These measures are used for selecting a unit of analysis. High synchronization for a unit pinpoints the level of the organization that exercises responsibility and control over the

Table 7

Synchronization Measures for Determining

the Unit of Analysis

| Unit of Analysis | Synchronization Definition[a] | Formula | Number |
|---|---|---|---|
| Companies ($C$) | $\frac{\sigma_C^2}{\sigma_C^2 + (q\sigma_S^2 + \sigma_E^2)/rsq}$ | $\frac{MS_C - MS_S}{MS_C}$ | (43) |
| Battalions ($B$) | $\frac{\sigma_B^2}{\sigma_B^2 + (q\sigma_S^2 + \sigma_E^2)/rsq}$ | $\frac{MS_B - MS_S}{MS_B + (c - 1)MS_S}$ | (44) |
| Brigades ($A$) | $\frac{\sigma_A^2}{\sigma_A^2 + (q\sigma_S^2 + \sigma_E^2)/rsq}$ | $\frac{MS_A - MS_S}{MS_A + (bc - 1)MS_S}$ | (45) |

[a]Subjects are considered random and items fixed. Formulas (44) and (45) differ from reliability formulas by an adjustment which makes the number of subjects within Brigades ($A$) and Battalions ($B$) hypothetically equal, for purposes of comparison, to the numbers within each company ($C$).

subject matter represented by the scale. These measures provide a way of directly comparing the extent of synchronization at each level of the hierarchy, $\underline{A}$, $\underline{B}$ and $\underline{C}$. At each level of hierarchy the number of subjects within the unit of analysis increases. Increases in subjects also increases reliability as measured by Formula 3. Reliability as measured by Formula 3 is again used as a synchronization measure, but only for the lowest level in the hierarchy--in this case for Companies ($\underline{C}$). The synchronization definitions and formulas for the higher levels of hierarchy $\underline{B}$ and $\underline{A}$ are adjusted statistically so that they have the same number of subjects within groups at the higher levels as was found at the lowest level $\underline{C}$. With this adjustment, the synchronization measures all become directly comparable. If a comparison of Battalion ( $\underline{B}$) and Brigade ( $\underline{A}$) synchronization is desired by itself, ignoring Companies ($\underline{C}$), the sample size adjustment can be made on Brigades, making Brigades equal in size to the level just below, Battalions, as follows:

$$\underline{S}_{\underline{B}} = (\underline{MS}_{\underline{B}} - \underline{MS}_{\underline{S}}) \ / \ \underline{MS}_{\underline{B}} \qquad (46)$$

$$\underline{S}_{\underline{A}} = (\underline{MS}_{\underline{A}} - \underline{MS}_{\underline{S}}) \ / \ \left(\underline{MS}_{\underline{A}} + (\underline{b} - 1)\ \underline{MS}_{\underline{S}}\right) \qquad (47)$$

where, $\underline{S}_{\underline{B}}$ equals synchronization for Battalions, and $\underline{S}_{\underline{A}}$ synchronization for Brigades.

## Significance of Difference Between Measures

With Formulas (43) to (45), the degree synchronization can be compared directly for each level of hierarchy, to determine the best unit of analysis. Finally, whether synchronization at one level is significatly greater than synchronization at another can be tested by forming appropriate quasi-$\underline{F}$ ratios as shown in Table 8. Each of the synchronization measures shares a common "error" term, $\underline{MS}_{\underline{S}}$, which is ignored when comparing relative sizes of synchronization measures, because it is held in common. Independent mean squares are needed for $\underline{F}$ ratios. Comparing synchronization can be accomplished by comparing the relative sizes of the "total" variance that has been adjusted for equal group sizes ignoring $\underline{MS}_{\underline{S}}$ for the reason stated. Company synchronization is compared to Battalion and Brigade synchronization in Formulas (48) and (49), and Battalion to Brigade in (50). For the latter comparison, Brigade size is adjusted to equal Battalion size in order to get a test with independent mean squares in the numerator and denominator of the $\underline{F}$ test. Power is greater for the test in Formula (50) than for the tests in (48) and (49).

When the hierarchical levels $\underline{A}$, $\underline{B}$ and $\underline{C}$ are confounded, individual scores may need to be weighted by Formula (41), to insure that each unit of analysis is weighted equally. The weights, when needed, will change as confounded hierarchical levels change. The coefficients $\underline{c}$ and $\underline{bc}$ in Formulas (44) and (45) are averages when the terms $\underline{A}$, $\underline{B}$, and $\underline{C}$ are confounded and weights are used. When different weights are applied at different hierarchical levels in a confounded

Table 8

Significance of Differences Between Synchronization Measures

| Comparison | Test Definition | $F$ Test[a] | Number |
|---|---|---|---|
| Companies ($C$) / Battalion ($B$) | $\dfrac{qrs\sigma_C^2 + q\sigma_S^2 + \sigma_E^2}{qrs\sigma_B^2 + q\sigma_S^2 + \sigma_E^2}$ | $\dfrac{c\ MS_C}{MS_B + (c - 1)\ MS_S}$ | (48) |
| Companies ($C$) / Brigade ($A$) | $\dfrac{qrs\sigma_C^2 + q\sigma_S^2 + \sigma_E^2}{qrs\sigma_A^2 + q\sigma_S^2 + \sigma_E^2}$ | $\dfrac{bc\ MS_C}{MS_A + (bc - 1)\ MS_S}$ | (49) |
| Battalion ($B$) / Brigade ($A$) | $\dfrac{cqrs\sigma_B^2 + q\sigma_S^2 + \sigma_E^2}{cqrs\sigma_A^2 + q\sigma_S^2 + \sigma_E^2}$ | $\dfrac{b\ MS_B}{MS_A + (b - 1)\ MS_S}$ | (50) |

Note. Formula (48) as written tests whether company synchronization is greater than battalion synchronization. The numerator and denominator can be reversed to test whether battalion synchronization is greatest.

[a]Degrees of freedom for quasi-$F$ tests are found by referring to Formula (24) in the text.

design, the mean squares in the numerator and denominator of the preceding tests are no longer independent, so that testing the significance of the difference between synchronization measures in this case should be used with caution.

### Removing Synchronization

When synchronization is found at more than one level of the hierarchy, the synchronization at the higher level can be partialed out using dummy regression, if desired. The existence of synchronization at each level can be tested by applying Formula (16) at each level of hierarchy to see if significant "true" variance can be identified at each level. The power of the test in Formula (16) is higher at higher levels. The number of degrees of freedom remaining after a higher-level group is partialed out may be reduced sharply as a result of removing synchronization. Removing synchronization from higher levels, however, would leave the researcher with results that could be unambiguously attributed to the lower-level unit and its leaders. Depending on hypotheses, this might be a desirable or an artificial result. It is possible, however, to statistically eliminate synchronization from higher levels when desired.

## Computational Requirements

There are two primary difficulties in computing the reliability and synchronization measures and tests given in this paper. The most serious difficulty is the computer core space required to compute a large split-plot analysis of variance design. All of the commonly used general analysis of variance packages, including SAS, RUMMAGE, BMD, MULTIVARIANCE, and IMSL, greatly exceed the core limitations of virtually all computers, for even modestly sized split-plot designs that involve even a moderate number of subjects. As the number of subjects in a split-plot design increase, factors that include subjects become huge. Commonly used analysis of variance packages attempt to store these huge factors in core. One exception is BMDP2V program, which does not require an unreasonable amount of core, but cannot compute the hierarchical portion of the design. Only one level in a hierarchy is possible. A general analysis of variance program capable of analyzing any design, was written to compute reliabilities for aggregated scores. The input data was organized by sorting to alleviate the cell storage problems. Multiple sorts are required for one run on a given model, but a large number of reliabilities can be computed during a single run.

The amount of computer CPU time taken to compute these reliabilities is a second problem. Most general analysis of variance packages create dummy variables to calculate either balanced or unbalanced designs, but in split-plot designs the number of dummy variables required is often huge, requiring large amounts of computer time. The general analysis of variance program that was written for computing reliabilities, uses the balanced algorithm given previously. The balanced algorithm is appropriate for unbalanced data when confounded components in an unbalanced design are allocated between true and error variance, as outlined previously. The algorithm was modified slightly in order to make the algebra appropriate in the unbalanced as well as the balanced case. Looking back at the steps required to get sums of squares, step (c) follows immediately after step (a) when applied to the unbalanced case. Degrees of freedom are obtained by getting the sum of the cells associated with main effects that are listed, instead of the product of the levels of the main effects listed, as given for the balanced case (see p. 4). The balanced algorithm in this program computes reliabilities much more rapidly than do programs that generate dummy variables. Multiple sorts on input data do, however, take some I-O ("wall clock") time, but this is required to alleviate the more serious core storage problems.[2]

## Summary

When research is conducted with intact organizations, groups rather than individuals are used frequently as the unit of analysis. One advantage of using groups as units is that, in this case, interaction within these groups can be studied. If groups are selected as the unit of analysis, what level of the organizational hierarchy should be selected for study? A statistical technique is suggested for selecting groups at the most appropriate level of the organizational hierarchy, at a level that actually controls and is responsible for the subject matter. This technique measures the extent of synchronization within groups at different levels of the hierarchy. The level selected for the unit should generally be the level with greatest synchronization.

After selecting an appropriate group unit of analysis, how should reliability be estimated? Survey variables consist of scores aggregated over both subjects within groups and survey items. The traditional methods of estimating reliability are either incomplete or inappropriate when applied to estimating the reliability of these aggregated scores. Using analysis of variance, appropriate reliability formulas were derived that depend on both the unit of analysis and survey sampling plan. In addition, significance tests for these reliabilities were given, as well as formulas to determine sample-size requirements from pretest data. A technique for estimating the reliability of record data, in the form of frequency counts within groups, is also given. Together, these statistical techniques provide improved methods for studying the operation of organizations.

[2] Information about the availability of this computer program may be obtained by writing the authors at Army Research Institute Field Unit, P.O. Box 5787, Presidio of Monterey, CA 93940. The program has been written so that it is easy to use with simple model input statements. Implementation on different computers could pose problems, depending on the extent to which the program is given continued attention and development by the authors.

## References


Bass, B. M., Valenzi, E. R., Farrow, D. L., & Solomon, R. J. Management styles associated with organizational, task, personal, and interpersonal contingencies. Journal of Applied Psychology, 1975, 60, 720-729.

Bowers, D. G. OD techniques and their results in 23 organizations. Journal of Applied Behavioral Science, 1973, 9, 21-43.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56 81-105.

Cronbach, L. J. & Furby, L. How should we measure "change"--Or should we? Psychological Bulletin, 1970, 74, 68-80.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Hart, R. J. Crime and punishment in the Army. Journal of Personality and Social Psychology, 1978, 36, 1456-1471.

Jones, A. P., & Jones, L. R. Psychological and organizational climate: Dimensions and relationships (Report No. 77-12). San Diego, Calif.: Naval Health Research Center, March 1977.

Jöreskog, D. S., & Sörbom, D. Advances in factor analysis and structural equation models. Cambridge, Mass: Abt Books, 1979.

Kenny, D. A. Cross-lagged panel correlation: A test for spuriousness. Psychological Bulletin, 1975, 82, 887-903.

Kerlinger, F. N., & Pedhazur, E. J. Multiple regression in behavioral research. New York: Holt, Rinehart & Winston, 1973.

McNemar, Q. Psychological Statistics (4th ed.). New York: Wiley, 1969.

Myers, J. L. Fundamentals of experimental design. Boston: Allyn & Bacon, 1966.

Passmore, W. A. The Michigan ICL study revisited: An alternate explanation of the results. Journal of Applied Behavioral Science, 1976, 12, 245-251.

Porras, J. I. The comparative impact of different OD techniques and intervention intensities. Journal of Applied Behavioral Science, 1979. 15, 156-178.

Satterthwaite, F. E. An approximate distribution of estimates of variance components. Biometrics, 1946, 2, 110-114.

Schneider, B. Organizational climate: Individual preferences and organizational realities. _Journal of Applied Psychology_, 1972, _56_, 211-217.

Schneider, B., & Bartlett, C. J. Individual differences and organizational climate II: Measurement of organizational climate by the multi-trait, multi-rater matrix. _Personnel Psychology_, 1970, _23_, 493-512.

Taylor, J. C., & Bowers, D. G. _Survey of organizations: A machine scored standardized questionnaire instrument_. Ann Arbor, Mich: University of Michigan, Institute for Social Research, 1972.

Torbert, W. R. Some questions on Bower's study of different OD techniques. _Journal of Applied Behavioral Science_, 1973, _9_, 668-671.

Winer, B. J. _Statistical principles in experimental design_ (2nd ed.). New York: McGraw-Hill, 1971.