Research Product 89-21

Questionnaire Construction Manual Annex

Questionnaires: Literature Survey and Bibliography

June 1989

Fort Hood Field Unit
Systems Research Laboratory

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

ERRATA AD-A213 255

FACSIMILE TRANSMITTAL HEADER SHEET

U.S. ARMY RESEARCH INSTITUTE
Scientific Information Office--Publications
5001 Eisenhower Avenue
Alexandria, VA 22333-5600
Comm # (703) 274-8029 or DSN 284-8029
FAX # (703) 617-0030

This fax is UNCLASSIFIED

From: U.S. ARI/E. Borg

Date: 06/10/96 **Time:**

Number of pages (including header) 1

To: Delores Campbell **Office symbol:**

Phone number: 767-9087 **Fax number:** 767-9070

Remarks: Reference: Research Product 89-21, " Questionnaire construction manual Annex. Questionnaires: Literature survey and bibliography" by Bettina A. Babbitt and Charles O. Nystrom, June 1989.

The following pages were intentionally left blank and not numbered in the report: iv, viii, 4, 6, 22, 30, 40, 44, 58, 62, 66, 82, 88, 92, 100, 108, 126, 128, 136, 164, 166, 176, 184, 232, 234.

ERRATA AD-A213 255

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS --
2a. SECURITY CLASSIFICATION AUTHORITY --	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE --	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Essex Corporation	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Research Product 89-21
6a. NAME OF PERFORMING ORGANIZATION Essex Corporation 6b. OFFICE SYMBOL (If applicable) --	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute Fort Hood Field Unit
6c. ADDRESS (City, State, and ZIP Code) 741 Lakefield Road, Suite B Westlake Village, CA 91361	7b. ADDRESS (City, State, and ZIP Code) HQ TCATA (PERI-SH) Fort Hood, TX 76544
8a. NAME OF FUNDING/SPONSORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences 8b. OFFICE SYMBOL (If applicable) --	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA903-83-C-0033
8c. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO.: 63739A; PROJECT NO.: 793; TASK NO.: 321; WORK UNIT ACCESSION NO.: A0
11. TITLE (Include Security Classification) Questionnaire Construction Manual Annex Questionnaires: Literature Survey and Bibliography	
12. PERSONAL AUTHOR(S) Babbitt, Bettina, A. (Essex Corporation), and Nystrom, Charles O. (ARI)	
13a. TYPE OF REPORT Final	13b. TIME COVERED FROM 83/05 TO 85/01
14. DATE OF REPORT (Year, Month, Day) 1989, June	15. PAGE COUNT 228
16. SUPPLEMENTARY NOTATION Contracting Officer's Representative, Charles A. Nystrom.	
17. COSATI CODES FIELD / GROUP / SUB-GROUP	18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number) Multiple-choice scales; Rank order scales; Bipolar scales; Paired-comparison scales; Semantic differential scales (Continued)

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

This report is an annex to the companion volume, "Questionnaire Construction Manual," published in 1985 by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI). It is designed to present summaries of the latest research findings related to developing questionnaires. Although both volumes were prepared primarily for personnel engaged in developing questionnaires for use in military tests and evaluations, the content is equally applicable to many nonmilitary areas.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Charles O. Nystrom	22b. TELEPHONE (Include Area Code) AV 738-9118 22c. OFFICE SYMBOL PERI-SH

ARI Research Product 89-21

18. SUBJECT TERMS (Continued)

Demographic characteristics
Continuous and circular scales
Behaviorally anchored rating scales
Questionnaire layout
Branching
Middle scale point
Response alternatives
Bibliography
Rating scales
Scale points
Item wording
Questionnaire construction

Research Product 89–21

Questionnaire Construction Manual Annex Questionnaires: Literature Survey and Bibliography

Bettina A. Babbitt
Essex Corp.

Charles O. Nystrom
U.S. Army Research Institute

ARI Field Unit at Fort Hood, Texas
George M. Gividen, Chief

Systems Research Laboratory
Robin L. Keesee, Director

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

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

June 1989

Army Project Number
2Q263739A793

Human Factors in Training and Operational Effectiveness

FOREWORD

This research was sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI), field Unit at Fort Hood, Texas, to develop a Questionnaire Construction Manual, Literature Survey, and Bibliography. The literature survey and bibliography present the latest research methods for developing questionnaires. The guidance contained will assist Army personnel in performing field tests and evaluations. Methods that are applicable to constructing questionnaires are described. The literature review and bibliography focus on content areas regarding scale categories, behavioral scales, design of questionnaire items, design of scale categories, interviewer and respondent characteristics, and questionnaire format. This research is a follow-on to the literature review of questionnaire and interview construction and administration conducted by Operations Research Associates in 1975 and edited and revised by the Army Research Institute in 1976.

EDGAR M. JOHNSON
Technical Director

ACKNOWLEDGMENTS

The preparation of this report was greatly facilitated by the generous assistance of several persons. A very special acknowledgment goes to Dr. Frederick A. Muckler, Essex Corporation, for his guidance and continuous support during all aspects of the preparation of this report.

The consultation and contribution of Mr. George M. Gividen, U.S. Army Research Institute for the Behavioral and Social Sciences (ARI), Commander William F. Moroney, Naval Air Development Center, and Dr. F. Thomas Eggemeir, Wright State University, are most gratefully acknowledged. Mr. Clarence A. Semple, Essex Corporation, contributed generously in editing. Mrs. Joan M. Funk, Essex Corporation, provided valuable technical assistance in preparing and editing the manuscript.

QUESTIONNAIRE CONSTRUCTION MANUAL ANNEX
QUESTIONNAIRES: LITERATURE SURVEY AND BIBLIOGRAPHY

EXECUTIVE SUMMARY

In 1975, Operations Research Associates (ORA) reviewed the literature on the construction and administration of questionnaires and interviews. Two publications resulted: a Questionnaire Construction Manual, which was revised/edited in 1976 to appear as an Army Research Institute special publication, P-77-1; and a Literature Survey and Bibliography Annex published as P-77-2. Also under contract to ARI, the Essex Corporation began in 1983 a survey of the literature for research done subsequent to ORA's cutoff date. The present volume is a sequel to P-77-2. It is intended for those concerned with questionnaire construction research from research design and developing scales to demographic characteristics of respondents.

Questionnaire construction research has not progressed evenly across professional fields. Sustained, programmatic research has hardly existed, whereas methodological considerations require a comprehensive series of experiments. In recent years, the computer has entered survey research. Its impact on construction, administration, and scoring is largely economic. Microprocessor, accessory, and software costs have continued to decline, and the efficiencies that result from computer use make its application very attractive.

Recommendations are provided for future research. Priorities are established for research topics as they relate to Operational Test and Evaluation performed by the Army Research Institute, Fort Hood, Texas. Topics covered are as follows: (1) scale development procedures and analysis; (2) procedural guides to item wording; (3) subjective workload assessment methods; (4) Automated Portable Test System; (5) cognitive complexity; (6) Behaviorally Anchored Rating Scales,; (7) item nonresponse, branching, and demographic characteristics; and (8) pictorial anchors.

QUESTIONNAIRE CONSTRUCTION MANUAL ANNEX
QUESTIONNAIRES: LITERATURE SURVEY AND BIBLIOGRAPHY

CONTENTS

CONTENTS (Continued)

QUESTIONNAIRE CONSTRUCTION MANUAL ANNEX
QUESTIONNAIRES: LITERATURE SURVEY AND BIBLIOGRAPHY

CHAPTER I
INTRODUCTION

In 1975, Operations Research Associates (ORA) reviewed the literature on questionnaire and interview construction, and administration research. They produced two products: a Questionnaire Construction Manual which was revised/edited in 1976, appearing as an Army Research Institute (ARI) special publication, P-77-1; and a Literature Survey and Bibliography volume published as P-77-2. Also under contract to ARI, Essex Corporation began in 1983 a search of the literature for research on questionnaires done subsequent to ORA's cut-off date. The present volume is a sequel to P-77-2. It is a companion volume that does not include the content of the previous work, although it does include the Table of Contents of P-77-2. This volume is, again, directed toward those who are tasked with questionnaire construction research ranging from research design, developing scales, through demographic characteristics of respondents.

To initiate the literature search, computer-assisted and manual searches were employed. The computer-assisted literature search accessed Dialindex across the following 20 data bases: ERIC, Educational Resources Information Center; NTIS, National Technical Information Services, U.S. Department of Commerce; SOCIAL SCISEARCH, Institute for Scientific Information; COMPENDEX, Engineering Information, Inc.; AIM/ARM, Center for Vocational Education; PSYCINFO, American Psychological Association; ABI/INFORM, Data Courier, Inc.; SCISEARCH, Institute for Scientific Information; COMPREHENSIVE DISSERTATION INDEX; SOCIOLOGICAL ABSTRACTS; MANAGEMENT CONTENTS; CONFERENCE PAPERS INDEX, Cambridge Scientific Abstracts; MENTAL HEALTH ABSTRACTS, National Clearinghouse for Mental Health Information, National Institute of Mental Health; ECONOMICS ABSTRACTS INTERNATIONAL, Dutch Ministry of Economic Affairs; U.S. POLITICAL SCIENCE DOCUMENTS, University of Pittsburgh Center for International Studies; HARVARD BUSINESS REVIEW, John Wiley & Sons, Inc.; HEALTH PLANNING AND ADMINISTRATION, U.S. National Library of Medicine; FIND/SVP REPORTS AND STUDIES INDEX; LC MARC, U.S. Library of Congress; BOOKS IN PRINT, R. R. Bowker.

Results from the Dialindex computer search suggested modification in the number of data bases to access. The 10 data bases which were used in the actual search and retrieval of citations were: ERIC, NTIS, SOCIAL SCISEARCH, COMPENDEX, PSYCINFO, ABI/INFORM, SOCIOLOGICAL ABSTRACTS MANAGEMENT CONTENTS, U.S. POLITICAL SCIENCE DOCUMENTS, and HEALTH PLANNING AND ADMINISTRATION. From the original computer-assisted literature search and the manual search, 16,816 citations were obtained, and 343 citations were identified as being potentially appropriate for questionnaire research. Subsequently, a supplemental computer-assisted Dialog search was run in the PSYCHINFO data base on the key word "Psychometrics." For the years 1976 through 1983, 2,415 citations were retrieved. Out of the 2,415 citations, 68 were under consideration for inclusion in the literature review. Subsequently, 178 citations were used in writing the sequel, although 463 citations on questionnaire methodology are found in the bibliography.

The content of the sequel was researched and written using the actual journal articles, reports, and books, and not the abstracts of the journal articles. Journal articles, reports, and books selected for inclusion in the bibliography were screened for their relevance to questionnaire construction. This sequel is designed to answer questions about the latest technical methods for developing questionnaires. These questionnaires are to assist Army personnel in performing field test evaluations. Methodological considerations which are relevant to constructing questionnaires, and could be generalized from other fields for military application, were used in conjunction with questionnaire construction research from the military. Relevant literature for questionnaire construction research from other fields included: political science, marketing, organizational management, human factors engineering, psychology, and education. Research on questionnaires was compared according to: description of subjects, number of subjects, number and type of experimental conditions, number of scale dimensions, number of scale points, response alternatives, hypotheses tested, results, scale reliability, and scale validity.

Each section in the sequel has been divided into four parts: (1) description of the content area, (2) examples of the content area, (3) comparison of studies, and (4) conclusions generated from the technical review. There are 27 different sections. Each section may be considered a stand-alone section. Each chapter subsection, II, 2.1-2.6; III, 3.1-3.4; IV, 4.1-4.5; V, 5.1-5.4; VI, 6.1-6.6; and VII, 7.1-7.2, for findings are restated in preference to directing the reader to another section.

The chapters contain related sections. Chapter II, Scale Categories, contains an overview for various multiple-choice scales that represent nominal, ordinal, and interval measurement. The assumptions underlying scale construction and developmental procedures are reviewed for bipolar, semantic differential, rank order, paired-comparison, continuous, and circular scales.

Chapter III, Behavioral Scales, consists of a wide variety of forms and methods to develop scales which have behavioral anchors. The developmental procedures for behavioral scales are addressed.

Chapter IV, Design of Questionnaire Items, expands upon contingencies involved in developing questionnaire items, such as the effectiveness of using positively and negatively worded items to create a balanced survey instrument. Other considerations include the number of items to use in a survey, and how many words to include in a question stem.

Chapter V, Design of Scale Categories, consists of the selection of number of scale points and type of response alternatives.

Chapter VI, Interviewer and Respondent Characteristics, views questionnaire construction from the standpoint of the impact on the target population, as well as on the interviewer, instead of the impact of the design of the instrument. Demographic characteristics which influence item responses are examined.

Chapter VII, Questionnaire Format, focuses on the physical structure of the questionnaire, the actual layout of the format, and the use of branching.

Chapter VIII, Future Research, is devoted to recommendations which will allow for systematic investigation of questionnaire construction for Army applications.

CHAPTER II

SCALE CATEGORIES

Well-known scales are reviewed in this chapter together with scale construction explanations based on the theoretical foundations developed by researchers, such as Thurstone, Likert, Guttman, and Osgood, Suci, and Tannenbaum. Examples of nominal, ordinal, and interval items, and response alternatives are provided. Scale category research is expanded upon in this section for bipolar, semantic differential, rank order, paired-comparison, continuous, and circular scales.

Since developmental procedures affect the statistical analysis obtained after scale administration, developmental procedures are important to ensure a quality scale. Guttman scales are suggested for applications with interval data. However, Guttman scales are more difficult to develop than other types of scales, and require greater development time. This constraint would be a hindrance in situations where Army personnel were participating in military field tests to assess equipment, training, organizations, and concepts, etc. due to the typical lack of developmental time. This constraint would apply to other scale categories to a lesser degree as well. The quality of any survey instrument depends on the quality of the developmental procedures.

In questionnaire construction, there have been no firm guidelines regarding when to use a checklist that forces a respondent into a dichotomous rating. It is suggested that checklists may be best applied in two types of situations. They are useful for rating observable job behaviors (this would be considered hard data), and for a presurvey to assist in developing refined items.

Even after items have been refined, there remains the issue of selecting response alternatives, and the question of what the midpoint is actually measuring (or for that matter, whether to use a midpoint). There is the possibility that in some instances subjects may be confounding scale dimensions with response alternatives. There has been evidence that response styles do exist, and the evidence has been conflicting. Apparently, minor violations in the development of response alternatives, and different types of response alternatives, have not jeopardized the reliability of instruments.

Overall research has not consistently shown one type of scale to be better than another. It has also been noted that the use of different types of statistics will generate different results with varying interpretations. Because of conflicting data, investigations have shifted to other aspects of questionnaire construction, such as: cognitive complexity of the respondent, and training respondents to use scales.

2.1 MULTIPLE-CHOICE SCALES

Description of Multiple-Choice Scales

In questionnaire construction there are two primary types of structured questions and response modes: (1) an open-ended question or (2) a multiple-choice question requiring a forced response. Researchers involved in the development of survey instruments usually use both types of questions. Open-ended questions serve well as preliminary screening devices for the development and refinement of multiple-choice questions (Orlich, 1978; Backstrom & Hurchur-Cesar, 1981).

While the world of questionnaires may be divided into these two categories, open-ended items require much less discussion because of their simplicity and limited role in questionnaires. Open-ended questions serve well when one is trying to determine what the relevant response alternatives to a question are. Thus, they enable the refinement of multiple-choice questions on the basis of the exploratory or pilot study administration (Orlich, 1978; Backstrom & Hurchur-Cesar, 1981). This is not to deny their utility on other occasions.

Multiple-choice items are preferred over open-ended items because of their potential for speed and objectivity in usage, provided that their development has involved sound procedures (Green, 1981). The number of response alternatives used with an item may range from 2 to over 20. The respondent may be directed to mark only one response choice, or may be allowed to select all response alternatives that seem appropriate to him/her. The choices may or may not be mutually exclusive (Orlich, 1978; Backstrom & Hurchur-Cesar, 1981).

Multiple-choice items represent measurement scales which are nominal, ordinal, or interval, and these scales indicate the rules for assigning numbers to the data so that the appropriate statistical analysis can be performed (Roscoe, 1975). Measurement scales for nominal items are non-numerical in their relationship. These items have mutually exclusive answers, and classify responses into categories (Roscoe, 1975; Orlich, 1978; Backstrom & Hurchur-Cesar, 1981).

Ordinal measurement scales have higher and lower categories, but the magnitude of the interval between responses is not specified. Unequal distances between intervals is always assumed, and the data is considered continuous when it is ranked (Roscoe, 1975). Ordinal measurement scales are common in surveys where respondents are required to rank items or to use a paired-comparison method (Backstrom & Hurchur-Cesar, 1981). This approach to scaling uses a Thurstone technique (Orlich, 1978). Usually, when 10 or more items are to be ranked, a Q Sort method should be used instead of a rank order scale.

Weighting scales for psychological distance or intensity can add exactness to a scale since it indicates how much difference there is among responses (Backstrom & Hurchur-Cesar, 1981). Interval measurement scales have equal intervals between the scale points (Roscoe, 1975), as well as retaining the characteristics of the previous scales.

Likert scales are the most widely used scales among researchers performing surveys (with the exception of market research surveys). Likert scales are usually composed of five or more response categories. The response categories for Likert scales are mutually exclusive and exhaustive (Backstrom & Hurchur-Cesar, 1981). Likert scales contain a statement of opinion followed by various levels of agreement or disagreement with that statement (Brannon, 1981). These rating scales are designed to present respondents with a statement, phrase, or word which describes their opinion or feeling. In addition to Likert scales, there are semantic differential scales, summed index scales, Guttman scales (Backstrom & Hurchur-Cesar, 1981), and Behaviorally Anchored Rating Scales (BARS). This list of scales is not meant to be inclusive.

Examples of Multiple-Choice Scales

In the design of a survey, researchers must decide whether to use an open question or a closed question with a multiple-choice format. The selection of a multiple-choice question automatically provides a fixed set of alternatives (Schuman & Presser, 1981).

Dichotomous item. Dichotomous items usually yield less variance than items with more response options. However, validity may suffer due to the lack of meaningful response alternatives (Brannon, 1981). In a test and evaluation of the Automated Shipboard Instruction and Management System, students aboard the U.S.S. Gridley were administered a questionnaire. Following is an illustration of several dichotomous items. This is a modified version of the Dollard, Dixon, and McCann (1980) Gridley Student Questionnaire.

	Yes	No
"Is this the first ship in which you have been required to qualify in General Damage Control PQS?"	___	___
"Are you familiar with the PQS booklet NAVEDTRA 43119-2A, 'Personnel Qualification Standard for Damage Control, Qualification Section 2, General Damage Control?'"	___	___
"Is your General Damage Control PQS progress charted in your divisional spaces?"	___	___
"Is the chart updated weekly?"	___	___

Shannon (1981a) used dichotomous questionnaire items for flight instruction primary training. The intent of these questions was to isolate recurring student problems during pre-solo training.

"1. Does this item represent a frequent error committed by the average student on all hops in primary training?"

"2. If the item is an error, is it critical?"

Multiple-choice -- fixed alternatives. Items which offer more than two alternatives are the most common types of items found in questionnaire construction. Sometimes an item (or a rating) has fixed alternatives where only one response alternative may be selected. An example of a fixed alternative with only one option is presented in modified form from the research of Bickley (1980). In this example, instructor pilots (IPs) were to select 1 of the 10 descriptors listed below after four maneuver repetitions by a student pilot.

Description of Maneuver for AH-1 Cobra Helicopter
Student Pilot Performance

(Select one descriptor)

___ "Demonstration by IP; no evaluation."

___ "IP immediately had to take back control of aircraft."

___ "Performance deteriorated until IP was finally obliged to take back control of aircraft."

___ "Student required considerable verbal assistance."

___ "Some parameters within course limits; verbal correction from IP required."

___ "Some verbal assistance required; less than one-half of parameters within course limits."

___ "Minimal verbal assistance; more than one-half parameters within course limits."

___ "Few parameters outside course limits; student corrected performance without coaching; still lacks good control touch."

___ "All parameters within course limits; work needed on control touch."

___ "Outstanding; no perceptible deviations from standards; SIP-level performance."

Multiple-choice -- select multiple alternatives. Some items are structured so that a respondent can mark all appropriate categories. In some instances, researchers construct a checklist to meet this objective. An example of an item with multiple alternatives was developed by Cicchinelli, Harmon, and Keller (1982). They constructed a checklist as part of an instructor questionnaire for a training simulation evaluation project.

> "What involvement have you had with the Denver Research Institute's evaluation of the simulated trainers? Please check any applicable statements."
>
> a. ___ "proctored the two-hour written test package"
> b. ___ "proctored the practical performance test"
> c. ___ "assisted with the design of the tests"
> d. ___ "was interviewed regarding my teaching methods and course material"
> e. ___ "had no involvement with the DRI evaluation program or development of materials."

Nominal item. Nominal response alternatives are typically mutually exclusive and often include precoded numbers used to identify the response alternative for data processing convenience. In the evaluation of observational skills, Block and Jouett (1978) had respondents rate a videotape of a clinical task performed by a respiratory therapist. Their rating form included nominal items and is modified for illustration below. Following is a nominal item developed by Block and Jouett to identify nonverbal interference factors during task performance.

> The form of nonverbal interference was:
> a) auditory
> b) visual

Ordinal item. Rankings can be used to order items in terms of importance or other dimensions. Below is an example of such a ranking task modified from the work of Hamel, Braby, Terrell, and Thomas (1983).

> "Format models on which learning aids are based present guidance on how to apply learning principles specific to a learning category." Rank the four statements below according to which statement you think is most important (one being most important, and four being least important):
>
> ___ Information is divided into small, easily learned blocks.
>
> ___ Illustrations present visual information such as the appearance of objects or signals, locations, and spatial relationships.
>
> ___ Distributed practice is provided through exercises, self-tests, and directions for remediation at appropriate points throughout the module.
>
> ___ Students are given immediate feedback on their responses within exercises.

Ordinal item -- paired-comparison. Backstrom and Hurchur-Cesar (1981) structured a paired-comparison item as a way to rank alternatives in a survey. Sources of information about federal involvement in a model city project was the topic area. A modified example of their paired-comparison method is presented here.

> Which do you generally find more reliable for obtaining information about the United States federal government involvement in city affairs?
>
> newspapers or radio
> radio or television
> television or newspapers

Ordinal item -- Q Sort. Ordinal measurements, where the paired-comparison items reach 10 items or more, are difficult to rank since ranking 10 pairs would require 45 different pairwise comparisons. A Q Sort technique can be applied in this type of situation. Moroney (1984) explains a sorting operation used with the Subjective Workload Assessment Technique (SWAT) which was developed by Shingledecker (1983).

> "SWAT is a two step process. Each individual scheduled to use SWAT participates in both a scale development phase and an event scoring phase. During the scale development phase, the person is asked to order a set of 27 cards from lowest workload to highest workload. The cards contain descriptions of levels of the three dimensions (i.e., time, effort, and stress). There are three levels of each of the three dimensions; therefore, all possible combinations result in 27 sets of descriptions. The individual's rankings of these sets of descriptions are then analyzed using conjoint measurement in order to find a mathematical model that describes the person's ordering. This model is then used to define scale values for workload from 0 for the lowest workload to 100 for the highest workload and 25 scale values in between. Thus, the scale is tailored to each individual's concept of how these factors combine to create the subjective impression of workload."

Ordinal -- Likert scale. Ordinal measurement scales do not assume equal distance between each scale point along a continuum of measurement. One of the common forms of ordinal scales is the Likert scale. Likert scales are usually composed of five or more response alternatives, each of which constitutes a point on the scale. Each question stem is followed by a scale, and the respondent is required to select only one scale point (response alternative) (Orlich, 1978).

In a survey of a training simulator evaluation project, Cicchinelli, Harmon, and Keller (1982) used this survey item with a 5-point scale ranging from 1, "disagree strongly," to 5, "agree strongly."

"From your general knowledge of and experience with simulated training, do you feel that simulated training:"

	Disagree Strongly				Agree Strongly
a. "is a good idea	1	2	3	4	5
b. can be more effective than actual equipment	1	2	3	4	5
c. can provide equivalent training with actual equipment	1	2	3	4	5
d. must be highly similar to actual equipment to be useful	1	2	3	4	5
e. can provide adequate training at a cost savings	1	2	3	4	5
f. allows for more complexity of training	1	2	3	4	5
g. is more reliable than actual equipment	1	2	3	4	5
h. teaches safety training better than actual equipment	1	2	3	4	5
i. provides more variety of training than actual equipment	1	2	3	4	5
j. is something you would use as an integral part of your teaching program	1	2	3	4	5
k. can replace actual equipment for 'hands-on' training"	1	2	3	4	5

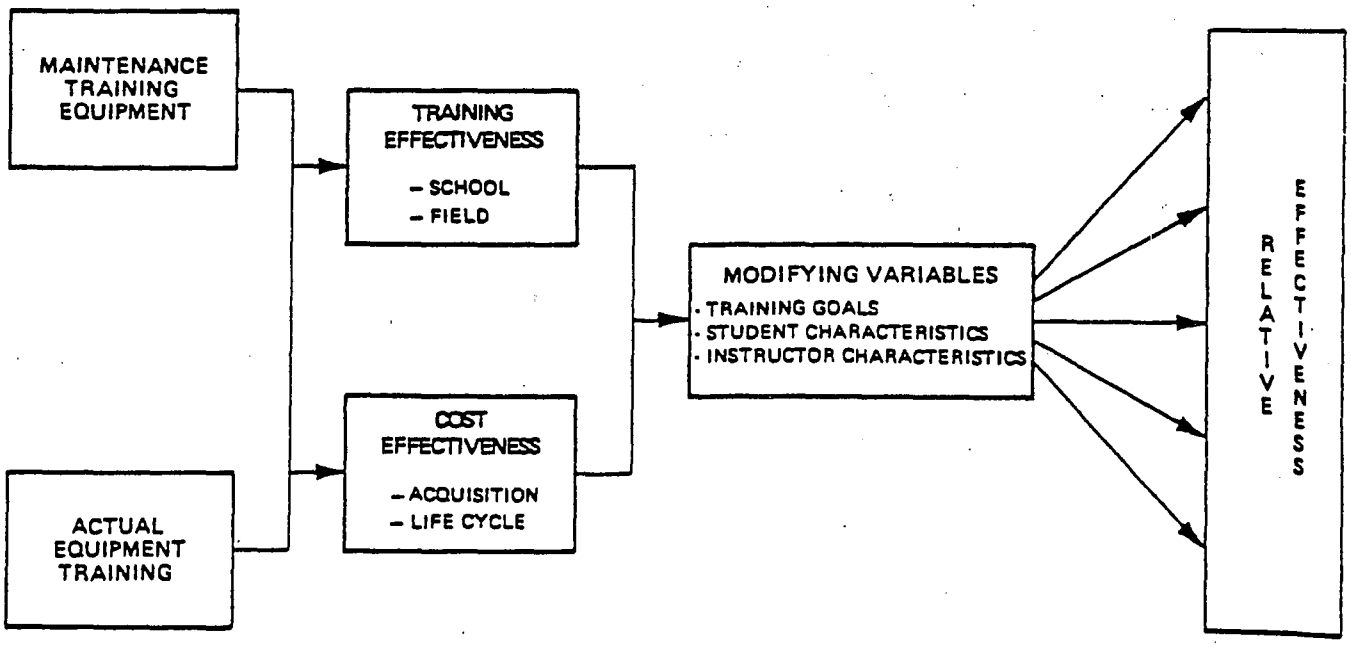
Interval item -- weighted. Equal distance between each scale point is assumed for interval scales. When constructing interval scales to measure the intensity of feeling, it is possible to design items where each response has a different weight assigned to it. The weights are used by analysts during analysis. The respondents are unaware what the weights are, and unaware that weights are being used. The assignment of weights would not be indicated on the questionnaire that respondents receive. An example of a survey item regarding public officials and disclosure of their sources of income is presented here.

> Would you say it is very important, fairly important, not too important, or not important at all that the Republican and Democratic vice-presidential candidates publicly disclose their private sources of income?

	(Weight)*
4 -- very important	(8)
3 -- fairly important	(7)
2 -- not too important	(3)
1 -- not important at all	(2)

*Weights not shown to respondents.

Interval item -- behaviorally anchored rating scale. Behaviorally Anchored Rating Scales (BARS) have traditionally been developed for performance appraisals. Wienclaw and Hines (1982) constructed BARS as a way to develop a valid tool to make decisions about the relative effectiveness of maintenance trainer equipment and actual equipment training. Their paradigm for determining relative effectiveness for the two training methods is presented here:



BARS were constructed to evaluate technicians' performance in field operations. Subject matter experts assisted in the development of BARS by identifying a series of critical incidents. Several hundred critical incidents were obtained. They described technician behavior on the job that differentiated between success and failure. The critical incidents were subsequently rated on a 7-point scale by instructors. Critical incidents which met statistical criteria were placed on a graphic rating scale and used to anchor the scale. Wienclaw and Hines (1982) identified seven specific dimensions by using the BARS technique. The seven dimensions are listed below:

1. "Safety: Behaviors which show that the technician understands and follows safety practices as specified in the technical data;"

2. "Thoroughness and Attention to Details: Behaviors which show that the technician is well prepared when he arrives on the job, carries out maintenance procedures completely and thoroughly, and recognizes and attends to symptoms of equipment damage or stress;"

3. "Use of Technical Data: Behaviors which show that the technician properly uses technical data in performance of maintenance functions;"

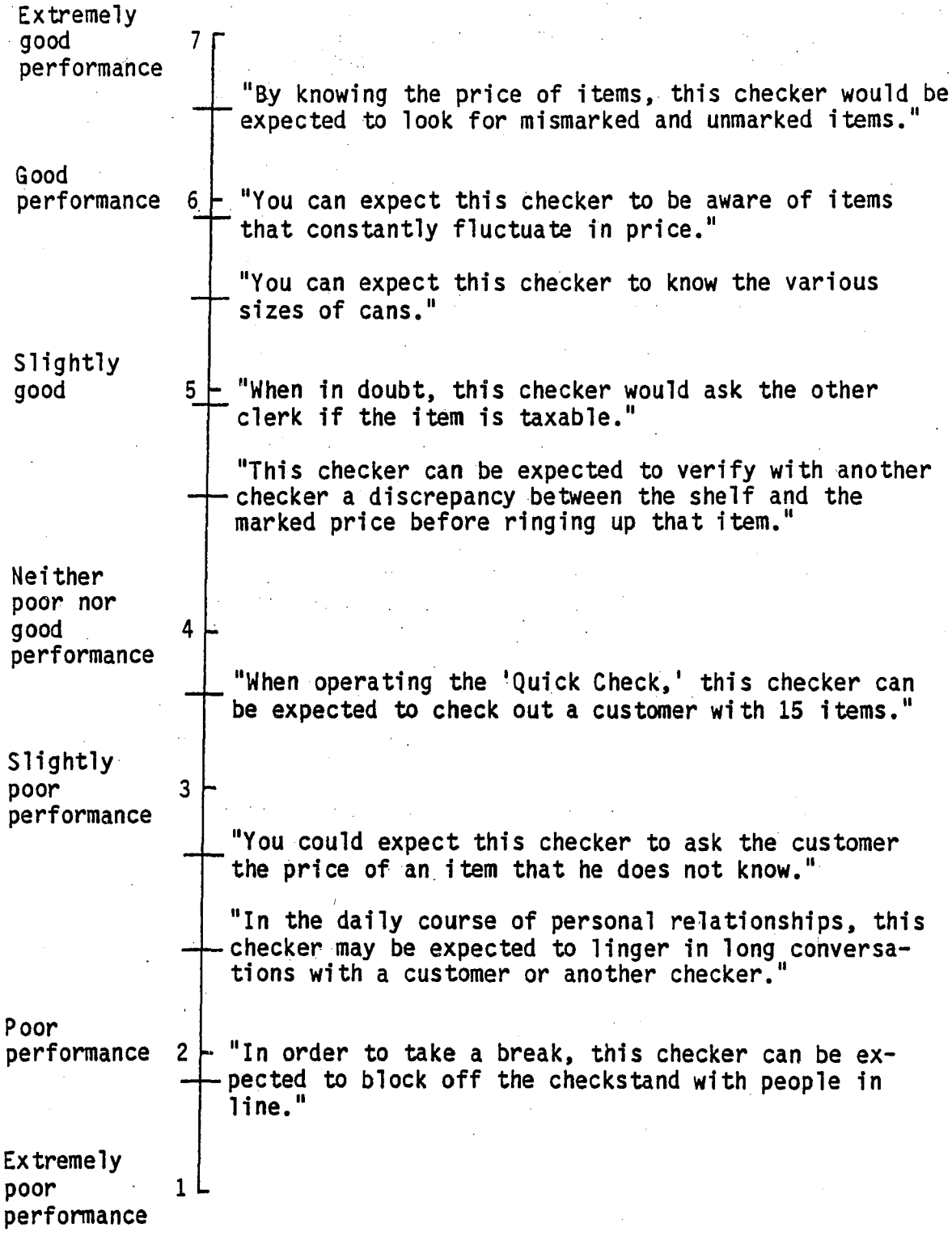
4. "System Understanding: Behaviors which show that the technician thoroughly understands system operation allowing him to recognize, diagnose, and correct problems not specifically covered in the Technical Orders and publications;"

5. "Understanding of Other Systems: Behaviors which show that the technician understands the systems that are interconnected with his specific system and can operate them in accordance with technical orders;"

6. "Mechanical Skills: Behaviors which show that the technician possesses specific mechanical skills acquired for even the most difficult maintenance problems; and"

7. "Attitude: Behaviors which show that the technician is concerned about properly completing each task efficiently and on time."

Kearney (1979) developed BARS to link appraisal to Management By Objectives (MBO) in an effort to reduce average customer check-out time. Illustrated here is their BARS for the performance dimension organization of the checkstand:



Interval item -- semantic differential. Interval measurement scales anchored by opposite adjectives on a bipolar scale, usually consisting of seven scale points, are known as semantic differential scales. Dickson and Albaum (1977) developed endpoint phrases by interviewing subjects to generate a representative sampling of descriptor phrases that could be used in their bipolar scale. To elicit their descriptors, they had their subjects use free association to label concepts, describe concepts in paragraph form, and develop paired sample bipolar endpoints with adjectives and with phrases. An example of the semantic differential scale is included below, and was developed by Dickson and Albaum for use in the study of retail images using adjectives and phrases as endpoints.

Bipolar Nominally Contrasting Adjectives and Phrases

crammed merchandise - well spaced merchandise
bright store - dull store
ads frequently seen by you - ads infrequently seen by you
low quality products - high quality products
well organized layout - unorganized layout
low prices - high prices
bad sales on products - good sales on products
unpleasant store to shop in - pleasant store to shop in
good store - bad store
inconvenient location - convenient location
low pressure salesmen - high pressure salesmen
big store - small store
bad buys on products - good buys on products
unattractive store - attractive store
unhelpful salesmen - helpful salesmen
good service - bad service
too few clerks - too many clerks
friendly personnel - unfriendly personnel
easy to return purchases - hard to return purchases
unlimited selection of products - limited selection of products
unreasonable prices for value - reasonable prices for value
messy - neat
spacious shopping - crowded shopping
attracts upper-class customers - attracts lower-class customers
dirty - clean
fast checkout - slow checkout
good displays - bad displays
hard to find items you want - easy to find items you want
bad specials - good specials

Interval item -- numerical scales. Interval items with numerical anchors have been used in human factors research at the Army Research Institute (ARI), Fort Hood. Listed below are examples of interval items developed by Dr. Charles Nystrom of ARI:

"Rate the effectiveness-ineffectiveness of the new weapon."
(Circle one of the numbers between the words.)

VERY EFFECTIVE							VERY INEFFECTIVE
+3	+2	+1	0	-1	-2	-3	

"Rate the effectiveness-ineffectiveness of the new weapon."
(Circle one of the numbers beneath the words.)

VERY EFFECTIVE	EFFECTIVE	IN BETWEEN	INEFFECTIVE	VERY INEFFECTIVE
+2	+1	0	-1	-2

"Rate the effectiveness-ineffectiveness of your performance of each of the tasks listed below."
(+2 = very effective, +1 = effective, 0 = in between,
-1 = ineffective, -2 = very ineffective, DK = don't know)

3.1	Starting the engine.	+2	+1	0	-1	-2	DK
3.2	Using the thermal sight.	+2	+1	0	-1	-2	DK
3.3	Erecting the flotation collar.	+2	+1	0	-1	-2	DK

Interval item -- summed index. Summed index scales use a series of agree and disagree statements to identify people who are typically conservative, authoritarian, liberal, etc. The summed number of agreements for an individual would determine differences among respondents on some characteristic.

Backstrom and Hurchur-Cesar (1981) used a summed index scale item to identify people who are typically conservative. A modified version of two items are illustrated:

All ethnic groups can live in harmony in the United States without changing our political system in any way.

Agree ___ Disagree ___

You can usually depend on a person more if they own their own home than if they rent.

Agree ___ Disagree ___

Interval item -- Guttman scale. Guttman scaling was developed as an alternative to Thurstone and Likert methods of attitude scaling, and is known as cumulative scaling and scalogram analysis. The underlying assumption of this interval scale is that subjects and items are both on a unidimensional continuum. McIver and Carmines (1981) provide an example of a perfect Guttman scale where it is possible to predict a perfect relationship between a scale score and a scale item (deviations from the model are always found in field applications).

Subjects	1	2	3	4	5	6	Scale Score
A	1	1	1	1	1	1	6
B	1	1	1	1	1	0	5
C	1	1	1	1	0	0	4
D	1	1	1	0	0	0	3
E	1	1	0	0	0	0	2
F	1	0	0	0	0	0	1
G	0	0	0	0	0	0	0

A Guttman scale to measure attitudes toward the Republican party was presented by Backstrom and Hurchur-Cesar (1981). The scale starts out with items that would be easy for Democrats to agree with and hard for Republicans to agree with. It continues on through the other end of the continuum so that only a rigid Grand Old Party (GOP) member could agree with the last statement.

Hard to Agree

"Generally speaking, the people of this country are better off electing a "bad" Republican president than a "good" Democratic president.

"Every Republican president has to try to reverse the unwise policies the Democrats enacted."

"Over the years, Republican presidents are more likely to act in the best interests of the country as a whole than are Democratic presidents."

"This country gets better government if the Republicans are in part of the time and the Democrats are in office part of the time."

Easy to Agree

"For all its faults, the two-party system of Republicans and Democrats is better than a one-party system."

Compound scale. Moroney (1984) presents three rating scales to the subject simultaneously. For each task, the respondent is to pick one rating only from the first scale, one only from the second scale, and one only from the third scale. The checklist joins together the three rating scales, and it joins together a multitude of tasks. Moroney included a

checklist completed by pilots which was developed by Helm and Donnell (1979) entitled Mission Operability Assessment Technique (MOAT). The following is a modified version of MOAT. This example presents a combining of stems and response alternative scales.

Listed below are a mission phase and a duty level and some of the tasks which are encompassed by them. Rate each task on the three scales by checking the appropriate line. Add any tasks which are not listed.	CRITICALITY OF TASK 1. Very small 2. Small 3. Moderate 4. Substantial 5. Very Substantial	PILOT WORKLOAD COMPENSATION/ INTERFERENCE 1. Poor 2. Fair 3. Good 4. Excellent	SUBSYSTEM TECHNICAL EFFECTIVE-NESS 1. Poor 2. Fair 3. Good 4. Excellent
MISSION PHASE: LAUNCH DUTY LEVEL: CONTROLLER			
TASKS:	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5
Control aircraft during takeoff rotation after catapult launch.	_ _ _ _ _	_ _ _ _ _	_ _ _ _ _
Control aircraft during configuration change. After including gear and flaps being raised.	_ _ _ _ _	_ _ _ _ _	_ _ _ _ _
Control aircraft during climbout.	_ _ _ _ _	_ _ _ _ _	_ _ _ _ _
Maintain appropriate internal/external scan of heads up and heads down instrument/displays during in-flight operations.	_ _ _ _ _	_ _ _ _ _	_ _ _ _ _
Monitor altimeter, air-speed, altitude, and heading on Heads-Up Dis-play (HUD) during launch.	_ _ _ _ _	_ _ _ _ _	_ _ _ _ _
Control aircraft during basic transitions from one flight altitude to another (climb, level-off, descent, turns).	_ _ _ _ _	_ _ _ _ _	_ _ _ _ _
ADDITIONAL TASKS	_ _ _ _ _	_ _ _ _ _	_ _ _ _ _

Developing survey items begins with a canvass of what questions ought to be asked. Following this is consideration of how to structure the response set for the respondent, and identifying the type of questionnaire layout. The statistical analysis selected will follow from the measurement scale displayed and response data obtained.

Comparisons of Multiple-Choice Scales

The research reviewed in this section on multiple-choice items was performed with samples containing college level students, with the exception of two studies representing Australian males (Ray, 1980) and computer-generated samples (Blower, 1981). No clear comparisons or conclusions were possible because of the different research designs used in comparing these items. For example, Blower measured a psychophysical procedure using a four-alternative multiple-choice task on a computer-generated sample. Deaton, Glasnapp, and Poggia (1980) measured the effects of frequency modifiers, item length, and statement direction.

Deaton, Glasnapp, and Poggia (1980) found a main effect for item positive and negative wording, and item length at the .05 level of significance. As item length increased, the average response rate moved toward the center of the response scale. Positively-worded items received higher mean responses than negatively-worded items.

Likert formats were used by Deaton, Glasnapp, and Poggio (1980), Ray (1980), and Bardo and Yeager (1982). Bardo and Yeager found that Likert formats were consistently affected by response style regardless of the number of scale points (4, 5, and 7). Ray compared measures of achievement motivation using a Likert behavior inventory format, forced-choice items, and a projective test. Behavior inventories, using a Likert format and forced-choice format, were both valid although a projective format was not.

Beltramini (1982) compared unipolar, bipolar, vertical, horizontal, and 5 and 10 scale point instruments to determine whether individual scale items were able to discriminate between two objects (black and white full page advertisements for a national fast-food restaurant) for the different formats used in this experiment. There were no significant interaction effects or main effects. Behavioral expectation scales were compared to checklists and graphic rating scales by Zedeck, Kafry, and Jacobs (1976). No conclusion as to format or scoring system superiority could be drawn from this research. Even so, different response formats and scoring systems led to different interpretations for performance appraisal scales.

Conclusions Regarding Multiple-Choice Scales

Results for multiple-choice scales are mixed. They do not lead to any concise conclusions. Replication of studies may be useful. Research that focuses on other variables, such as training, cognitive style, scale developmental procedures, and other variables, other than format variations, may bring about more fruitful lines of research.

No one multiple-choice type of format can be recommended. Likert scales appear to be statistically superior to Thurstone scales, and Guttman scales are statistically superior to Likert scales (McIver & Carmines, 1981). Guttman scales should be used with interval measurement only, and are the most difficult to develop. Guttman scales have been used to measure psychophysical phenomena (Blower, 1981; Jesteadt, 1980) and attitude survey items (McIver & Carmines, 1981; Backstrom & Hurchur-Cesar, 1981). Guttman scaling theory is used in the expanding field of adaptive testing.

2.2 BIPOLAR SCALES

Description of Bipolar Scales

Bipolar scales are usually associated with semantic differential scales (Klockars, King, & King, 1981). Bipolar scales are traditionally anchored by verbal labels at the endpoints. It is assumed that the scales have bipolarity since they are usually anchored by adjectives which are antonyms (Mann, Phillips, & Thompson, 1979). As semantic differential scales, they have been used extensively in marketing research. In addition, Army Research Institute (ARI), Fort Hood, TRADOC Combined Arms Test Activity (TCATA), and Operational Test and Evaluation Agency (OTEA) have used bipolar scales almost exclusively in their human factors assessments of Army systems, organizations, and training. Bipolar scales have been extensively used for self-description in personality assessment, although there have been other applications for these scales. (Army Research Institute, Fort Hood, Texas has been using bipolar scales, but not in semantic differential format.)

The semantic space between the bipolar anchors theoretically have a three-factor structure: evaluation, potency, and activity, which was introduced by Osgood, Suci, and Tannenbaum (1957). The three-factor structure introduced by Osgood et al. has been found to be present when measuring personality traits and attitudes. The application of bipolar scales for human factors assessments of Army systems cannot be assumed to have the same underlying three dimensions. Evaluation would be the primary dimension used in the assessments of Army systems. Included in the evaluation dimension are the components of evaluation of human factors, such as: effectiveness (+,-), adequacy (+,-), satisfactoriness (+,-), timeliness (+,-), and accuracy (+,-). Mann, Phillips, and Thompson (1979) mentioned that there has been the assumption that a line anchored by the polar terms has opposite meaning and equal distance between the two symmetrical poles. This assumption has not been totally supported by research. It does not account for the center of the scale (zero point), without which one cannot tell where one meaning leaves off and its opposite starts. It is assumed that the distance from the midpoint to Pole A is equal and opposite the distance from the midpoint to Pole B.

Construction of bipolar scales embedded in the semantic differential frequently uses a series of seven intervals along the scale line. Some researchers use other numbers of scale intervals, such as: 5 and 11 (Johnson, 1981; Eiser & Osmon, 1978; Klockars, King, & King, 1981). The bipolar scales are often anchored by four adjective trait terms. The scales are divided into subsets so that each adjective is used as an endpoint only once in each subset. Klockars et al. provide an example of bipolar endpoints using Peabody's 1967 four adjective trait terms: Cautious-Bold, Rash-Timid, Cautious-Rash, Bold-Timid. Other variations for the identification of endpoints on bipolar scales have also been developed. For example, quasi-polar scales were developed by using partial antonyms of undetermined functional antonymity (Vidali, 1976). Beard (1979) used bipolar scales with pictorial anchors, and Dickson and Albaum (1977) used phrases as endpoints.

Examples of Bipolar Scales

Dolch (1980) compared numerical bipolar scales and adverb bipolar scales on a semantic differential to measure students' feelings to evaluate a text for introductory sociology. An example of one of his numerical scales is as follows:

> "Below is a series of adjectives which might be used to describe the Caplow text. Circle the number which best expresses how you feel.
>
> Important 3 2 1 0 1 2 3 Unimportant
>
> If you feel the book is really important, circle 3."

The adverb scale varied from the numerical scale by placing adverbs at the scale points instead of numbers. Subjects using the adverb scales were requested to circle the adverb that best expressed their feelings.

Bipolar scales are widely used with the semantic differential technique. Researchers have selected bipolar scales using the semantic differential that appeared to be appropriate to measure various content areas. When new bipolar scales are developed, they need to be tested for their psychometric properties, the bipolarity of the endpoints, and for the underlying assumptions of the semantic differential.

In the bipolar items that the ARI, Fort Hood, Texas uses, the researchers started out using scale lines, but have reduced the frequency of such use greatly. The scales use a horizontal layout; a scale line could have been penned in if it was worth the effort. ARI researchers also use the same response alternatives in a vertical format, both with and without numerical values preceding the positive and negative response alternatives, and a "0" in front of the midpoint response alternative. It's probably somewhat less obvious that the researchers are suggesting a scale when using the vertical format, but they are. The prime example of a scale is a ruler. Most rulers are unipolar and have three elements: the numbers, the tick marks, and the line. ARI, Fort Hood, Texas has gotten away from using the tick marks and the line, but still uses the numbers. The ARI researchers use a variation influenced by the scales one finds (or used to) in an algebra book. That is, they have a conceptual line with a "0" centered along it; negative numbers running in one direction, and positive numbers running in the opposite direction (left or right makes no difference to the scale, although in algebra the negative numbers run to the left or downward). When unlined scales are used with word anchors at the ends and intermediate points with numbers beneath the words, the numbers may not always be equally spaced. There is no deliberate distortion sought or deviation from the appearance of equal spacing of the response alternatives along the conceptual line.

The Nystrom Number Scale is based on an algebraic number scale. In an earlier version of this scale, antonyms were placed above the numbers rather than at the two ends of the string of numbers. The concept was to label the two directions without overly influencing or anchoring the meaning of the end numbers. The result might be that respondents would make

more frequent use of the extreme numbers. Below is an example of such a scale:

Rate the effectiveness-ineffectiveness of the new M1E1 main gun: (Circle only one of the numbers below to show your rating.)

EFFECTIVENESS				INEFFECTIVENESS		
+3	+2	+1	0	-1	-2	-3

The following format has been widely used by ARI, Fort Hood:

Important +3 +2 +1 0 -1 -2 -3 Unimportant

Approximately 5% of the respondents tended to circle the end words rather than circling the numbers. (This may have been due to the limited amount of guidance for respondents on how to use the scale.) To avoid this problem in the future, TCATA selected a modified version of the above scales. The revised scale includes five sets of word anchors with an algebraic number under each, as shown in Section 2.1, Interval item -- numerical scales.

Comparison of Bipolar Scales

The subjects reported in the literature reviewed for bipolar scales consisted almost exclusively of students ranging from eighth grade through graduate school (as well as their wives) (Eiser & Osmon, 1978; Dickson & Albaum, 1977). In one sample, subjects were identified as male readers of Horizons USA who resided in Great Britain, Italy, Phillipines, and Venezuela (Johnson, 1981). The number of scale points ranged from 5 through 11. Endpoints for the bipolar scales varied across studies, although adjectives which were antonyms were used most frequently. Beard (1979) anchored the endpoints with pictures, Vidali (1976) anchored endpoints with bipolar and quasi-polar adjectives and adverbs, while Dickson and Albaum (1977) anchored endpoints with adjectives and phrases. ARI, Fort Hood, has anchored endpoints with various bipolar formats that have included only antonyms, only numbers, and both antonyms and numbers.

One of the main concerns when anchoring bipolar scales is the tendency to consistently use a response style which favors a positive or negative anchor (Johnson, 1981). In the case of trait assessment, there is a tendency to use a socially desirable response style (Klockars, King, & King, 1981; Klockars, 1979; Eiser & Osmon, 1978).

In a cross-cultural study regarding the order of presentation of stimulus words (positive or negative anchors for the bipolar scale), there were no clear differences in the means for the ratings on eleven dimensions. This resulted from placing the positive or negative response first on a bipolar scale (Johnson, 1981). Overall, the effects of response style were negligible, but there is evidence that response style may vary from country to country. Johnson described response style for the cross-cultural study as a consistent tendency by respondents to answer survey items positively or negatively dependent on stimulus words. Two questionnaires were developed for this study. One of the questionnaires had the

positive stimulus words presented first. The other questionnaire had the negative stimulus words presented first. (This was on a semantic differential scale with 11 intervals from 0-10.) Johnson performed sign tests to determine the significance of differences between the means within each country. The sign test was significant at the .05 level for a response style in the Philippines, and in Italy. These results indicate that there is a tendency by respondents in the Philippines to use a positive response style, and by respondents in Italy to use a negative response style. Respondents from Britain and Venezuela had no response style related to the order of presentation for the two questionnaires. Johnson suggested that bipolar adjective scales have not usually been affected by the placement of stimulus words across national studies, but that cross-cultural studies may require taking response style into consideration for homogeneous groups, especially when the situation is ambiguous and/or unstructured (see Section 4.5, Balanced Items).

Klockars, King, and King (1981) and Klockars (1979) explored bipolar scales for social desirability responses. Klockars et al. used sets of bipolar scales in the semantic differential format to measure the subjects' (psychology students) self-description for 13 different personality traits. Scales were constructed so that they had both positive (desirable) endpoints, both negative (undesirable) endpoints, and a combination of one positive (desirable) endpoint and one negative (undesirable) endpoint. It was assumed that the underlying structure for the connotative meaning is composed of evaluation, potency, and activity (see Section 2.3, Semantic Differential Scales). They explored the dimensionality of the bipolar scales used in self-description for personality assessment. It has been argued that there is a social desirability response (related to the evaluation portion of the underlying structure of the semantic differential), and that it may confound the response style on personality instruments. They investigated whether the social desirability responses were predominant in self-ratings. It was determined that the scores were internally consistent and were not correlated with social desirability. They were not able to obtain evidence to support a social desirability response tendency.

Klockars (1979) felt that when both endpoints on a bipolar scale were anchored with verbal labels that there was the possibility of confounding ratings with trait (for personality scales), and social desirability responses. This research was similar to that reported above by Klockars, King, and King (1981). The scales that were constructed by Klockars (1979) were all trait scales. The results were confounded whether the stem was a desirable or undesirable adjective. These findings were significant at the .05 level indicating that subjects systematically rate scales so that the desirability dimension is confounded with the trait dimension. Klockars (1979) compared the strength of the social desirability effect when the stem words were undesirable. The level of significance obtained was .05. When a socially undesirable adjective is presented, there is a propensity for subjects to select an adjective which is opposite in desirability. The results obtained by Klockars (1979) are in conflict with the Klockars, King, and King (1981) findings.

In a study performed by Eiser and Osmon (1978), bipolar scales were constructed. Half of the scales consisted of endpoints anchored by positive labels at both ends of the scale. The other half of the scales were

anchored by negative labels at both endpoints of the scales. They hypothesized that when a scale was anchored at both ends by negative labels, the responses obtained should represent a wider perspective and have less polarized ratings. This should be irrespective of the attitudes of the respondents. They also hypothesized that scales which were anchored at both endpoints by positive labels would have more polarized ratings. Their findings indicated that subjects gave more polarized ratings at the .001 level of significance for scales with endpoints which were both positively labeled, as well as for scales with both endpoints negatively labeled. The middle portions of the scales may have been perceived as being neutral when both endpoints were labeled positive or negative. They indicated that raters tended to give positive responses to items they agreed with, and negative responses to items they disagreed with. For items respondents agreed with, they tried to avoid giving a negative response. Respondents tried to avoid giving positive ratings to items they disagreed with. These researchers intended that the scales used in this study be symmetrical in terms of grammatical form and evaluation. They determined that the effects of the response language (positive or negative) used for endpoints on a bipolar scale can influence the response, independent of the subjects' attitudes.

In situations where researchers are using bipolar scales as a vehicle for determining the influence of positive and negative anchors, they may at times be violating the theories relevant to the underlying structure of their scales. For example, it is assumed that bipolar scales are anchored by adjectives which are antonyms (Mann, Phillips, & Thompson, 1979). When a scale is anchored by endpoints which have labels that are both positive or both negative, the researchers have violated the assumption of bipolarity. More research may be required on bipolarity because of: violation of the basic assumption of bipolarity, conflicting research results, and paucity of research on the topic. There is not clear evidence to substantiate the effects of: the influence of positive or negative endpoints on bipolar scales, and the effects of social desirability responses and their confounding with other variables (e.g., the evaluation factor found in semantic differential scales).

Other research has focused on bipolar endpoints which differ in other ways than positive and negative anchors. Dolch (1980) compared bipolar scales anchored with numerical or adverb responses. The correlation between the two scales was -.929. On the surface, it did not appear to matter which type of endpoints were used. A factor analysis of the two scales revealed markedly different factor structures which indicates that the two scales were not measuring meaning in the same way.

The differences between scales anchored by bipolar or quasi-polar adjectives and the effects of concept interaction were examined by Vidali (1976). Quasi-polar scales contained anchors that were considered to be only partially antonymous. The effects of concept-scale interaction on reliability did not impair the reliability of the scales. There was an interaction effect when the scale was used with certain concepts identified as "unstable." The inadvertent use of mismatching scales (by using antonyms with partial antonyms) did not appear to jeopardize the reliability of the scale.

In an unusual approach in the development of bipolar scales, Beard (1979) anchored endpoints with pictures instead of by the common verbal anchoring. Beard anchored bipolar scales with pictorial anchors through the use of color slides and rating forms with replicas of the slides. The pictorial anchors were not verifiable as antonyms. There may be an application for this type of measurement in human factors research on equipment designs, and for respondents who are limited by language facility yet have cognitive strengths for spatial differentiation. New developmental techniques and methods would have to be established for group administration.

The studies cited in this research in bipolarity have been diverse in the variables measured, the analyses applied, and the results obtained. Applicability is limited to students and survey application to academic environments.

Conclusions Regarding Bipolar Scales

The assumption of bipolarity for scaling purposes is that Pole A to Pole B is 180^{0}. In application, scale bipolarity may be approximate. Scales do not always meet the criterion of bipolarity mentioned above. The variables that have affected bipolar scales have been: the differences among how respondents rate the scales, the issue of the relevancy of the scale to the respondent, and the assumptions about the psychometric qualities of the scale as developed by Osgood, Suci, and Tannenbaum (1957).

Conformation in studies was not found for social desirability responses, and first presentatiton of endpoints for positive or negative anchors. Subjects may be confounding trait dimensions with response anchors. It is possible that some individuals may make greater use of the extreme categories at the ends of the scale because they are influenced by the descriptive anchors (Johnson, 1981; Eiser & Osmon, 1978; Klockars, 1979; Klockars, King, & King, 1981). There is no clear evidence to support the existence of response style associated with the order of positive or negative anchors.

The meaning of the midpoint is also of concern for bipolar scales, behavioral observation scales, behavioral expectation scales, behaviorally anchored rating scales, etc. There is some question about what the midpoint is actually measuring (neutrality, ambivalence, or irrelevance). According to Mann, Phillips, and Thompson (1979), respondents may include an irrelevance response separate from the scale midpoint, such as the "Don't Know" category. Variations in instrument format and instruction did not alter the scale dimension. These bipolar scales did not provide a separate "Don't Know" category (see Section 5.2, "Don't Know" Category, and Section 5.4, Middle Scale Point Position).

Bipolar scales have had many applications. For example, Dickson and Albaum (1977) were able to successfully develop a marketing survey on retail store images for supermarkets, department stores, shoe stores, and discount stores using a semantic differential format. This indicates that survey researchers may want to explore the use of the semantic differential when developing new bipolar instruments.

Bipolar scales have proven to be psychometrically sound when using the semantic differential format. Manipulation of the anchors for type of anchor or presentation of positive/negative anchors does not appear to greatly affect the results. Research on response sets has not been consistent so that a trend cannot be cited.

2.3 SEMANTIC DIFFERENTIAL SCALES

Description of Semantic Differential Scales

Semantic differential scales were developed by Osgood, Suci, and Tannenbaum in 1957 (Klockars, King, & King, 1981; Downs, 1978; Maul & Pargman, 1975). A concept or descriptive term is presented to the respondent (Maul & Pargman, 1975). These scales are usually anchored by adjectives with opposite meanings at the endpoints (Backstrom & Hurchur-Cesar, 1981; and Klockars, King, & King, 1981). Semantic differential scales almost always have a horizontal bipolar format with seven scale points (Church, 1983). Some scales have been known to have fewer scale points (Albaum, Best, & Hawkins, 1981; and Vidali, 1976).

The underlying assumption of the semantic differential scale is that there are three major factors for the measurement of concept in the semantic space (Klockars, King, & King, 1981; Malhotra, 1981; Dziuban & Shirkey, 1980; and Maul & Pargman, 1975). The three major factors accounted for in the semantic space are: Evaluation, Potency, and Activity (EPA). The evaluation factor is responsible for the greatest amount of variance (Klockars, King, & King, 1981). The dominant evaluative factor indicates a good-bad perception by the respondent. Perception of the potency factor is related to a strong-weak relationship, and the activity factor indicates a perception of fast-slow (Maul & Pargman, 1975).

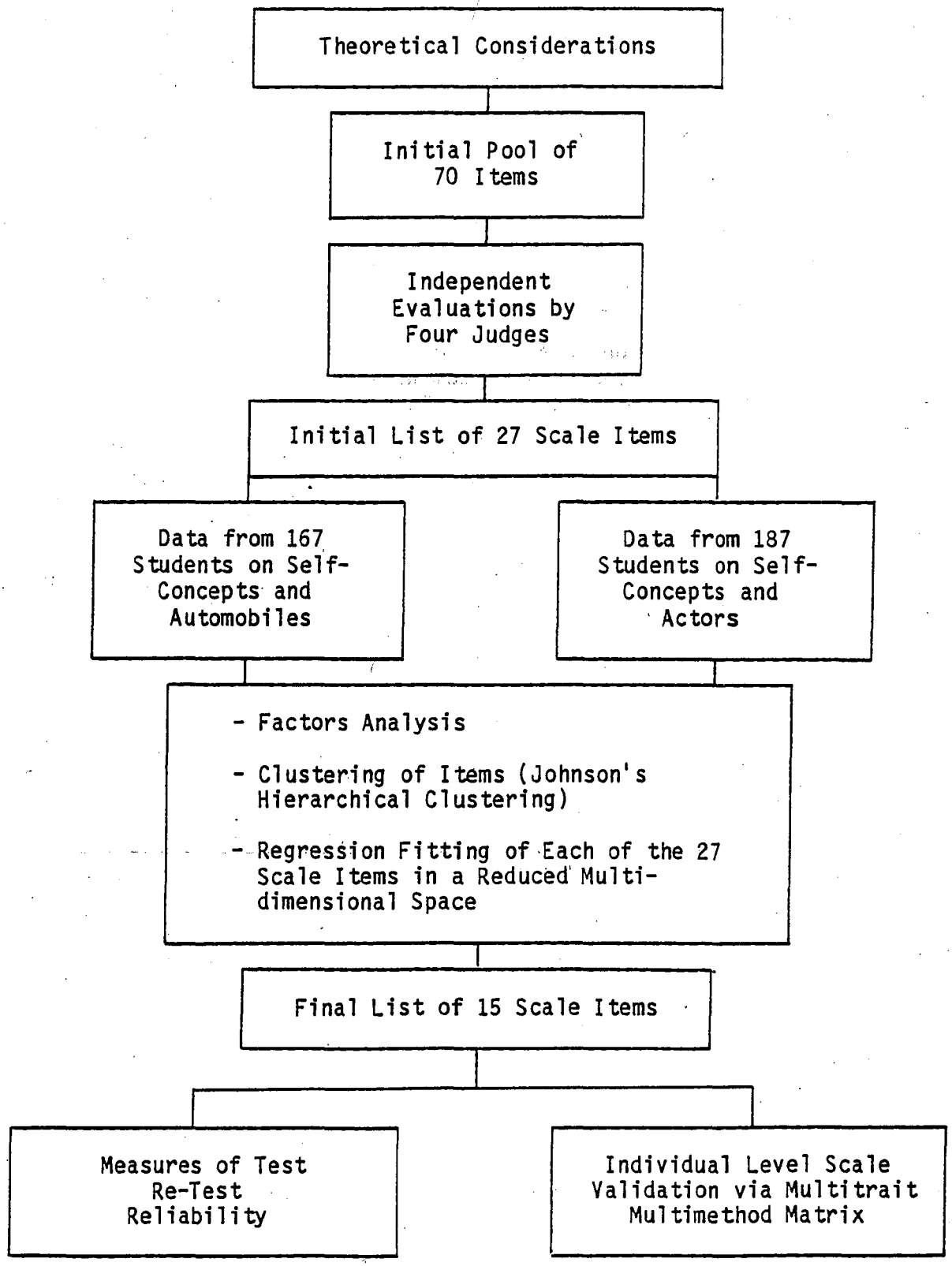
Semantic differential scales have been used by researchers in multiple fields, such as marketing, education, and psychotherapy. Marketing researchers have used this instrument extensively (Malhotra, 1981; Downs, 1978). These scales measure attitudes and opinions (Church, 1983). Attitudes may include unconscious or nonverbalized avoidance tendencies. Opinions are restricted to verbalized attitudes. The concepts of attitude and opinion are closely aligned and not always overlapping (Kiesler, Collins, & Miller, 1969).

The selection of anchors for endpoints has been accomplished in various ways. One approach has been to select anchors through free association of subjects for concepts, and through the use of dictionaries and thesauruses. After a pool of items has been compiled, agreement by judges facilitates a reduction in the number of items. Factor analysis and cluster analysis can also be used to determine which items load on the same factor, and which items tend to cluster together. This allows for further reduction in the number of items (Malhotra, 1981). The selection of items allows for instruments which are individually designed for specific research projects (Dziuban & Shirkey, 1980). This is an important aspect of scale development for the semantic differential since the meaning of an item and its relationship to other items will change depending on what concept is being assessed (Dickson & Albaum, 1977).

Examples of Semantic Differential Scales

Semantic differential scales have been developed with different endpoint anchors, such as adjectives, adverbs, and phrases. These scales have

been designed to measure various concepts, attitudes, traits, etc. In research performed by Malhotra (1981), a scale was constructed to measure specific concepts related to automobiles and actors. Data was also obtained on three measures of self-concept: ideal self, actual self, and social self. Malhotra formulated the flow of the developmental procedure for the semantic differential used to identify items. Following is Malhotra's flowchart of the scale development procedure:



Fredericksen, Jensen, and Beaton (1972) investigated adjectives that they hypothesized would be relevant to organizational climate or to a subject's reaction to the organizational climate. Following is an example of their semantic differential scale:

"Encircle the number that best describes the subject and/or his behavior."

7.	Compulsive	9	8	7	6	5	4	3	2	1	0	Noncompulsive
8.	Flexible	9	8	7	6	5	4	3	2	1	0	Rigid
9.	Global concerns	9	8	7	6	5	4	3	2	1	0	Specific Concerns
10.	Ordinary	9	8	7	6	5	4	3	2	1	0	Creative
11.	Authoritarian	9	8	7	6	5	4	3	2	1	0	Democratic
12.	Careful	9	8	7	6	5	4	3	2	1	0	Careless
13.	Satisfied	9	8	7	6	5	4	3	2	1	0	Disgruntled
14.	Complaisant	9	8	7	6	5	4	3	2	1	0	Rebellious

Downs (1978) developed three versions of the semantic differential. Version A is an upgraded semantic differential. It was orginally developed by Hughes (1975) in the hope of reducing anchoring problems, halo error, and the number of items. The subjects were students who were requested to rate alternative living quarters that they were familiar with. Version A requested the respondent to rate 10 residences on a scale of 1 through 7, from "least preferred" to "most preferred."

"Rate the 10 residences in terms of how much you would like to live in each."

Least Preferred	C						Most Preferred
	/ MW /	/ D /	F /	JL /	V /	PY /	
	1	2 3	4	5	6	7	

C = Conway
D = Dormitory (coed)
W = Woodshire
M = Male/Female Dorm
Y = Yancey Motel
F = Fraternity/Sorority
J = James Blair Terrace
P = Parkway
L = Ludwell
V = Village

Version B is a semantic differential scale format that is frequently used in marketing questionnaires. Downs' example of the marketing research approach is shown below:

"Rate the 10 residences in terms of how much you would like to live in each. Merely circle the number that reflects your preference for each residence."

	Extremely Low Preference						Extremely High Preference
Conway Apartments	1	(2)	3	4	5	6	7
Dormitory (coed)	(1)	2	3	4	5	6	7
Woodshire Apartments	1	2	3	4	5	(6)	7
Male/Female Dormitory	1	2	3	4	(5)	6	7
Yancey Motel	1	(2)	3	4	5	6	7
Fraternity/Sorority House	1	2	3	(4)	5	6	7
James Blair Terrace	1	2	3	4	(5)	6	7
Parkway Apartments	1	2	3	4	(5)	6	7
Ludwell	1	2	3	4	5	6	(7)
Village Apartments	1	2	3	4	5	(6)	7

The last version that Downs (1978) devised consisted of a format more along the lines of the traditional semantic differential scale (in this instance, anchored by adjective phrases).

Comparisons of Semantic Differential Scales

Research on the semantic differential scales has been difficult to compare since instruments are not constructed consistently. They do not use the same number of scale points or similar types of anchors. For example, Dickson and Albaum (1977) anchored their semantic differential scale with phrases and adjectives, Dolch (1980) anchored semantic differential scales with adverbs and numbers, and Vidali (1976) anchored scales with what was termed bipolar and quasi-polar adjectives.

Of particular concern is the structure of concepts in the semantic space where most of the variance has been accounted for by the concepts of evaluation, activity, and potency (Dziuban & Shirkey, 1980). There have been different approaches in measuring the semantic space. Mann, Phillips, and Thompson (1979) studied the issue of the bipolarity of the semantic space. They found that the scale x concept x person interaction was responsible for a greater part of the variance than a concept x scale interaction. Individual differences influenced the three-way interactions. This affects the interpretation of the three-dimensional (evaluation, activity, potency) semantic space. The three-dimensional semantic space is not found to be descriptive of all subjects. However, overall, the three-dimensional structure of the semantic space is robust when all subjects are taken into account since variations in format and instructions don't appear to change the three-dimensional structure for the sample as a whole.

Psychometric adequacy for the concept structure of the semantic space was examined by Dziuban and Shirkey (1980) using the Measure of Sampling

Adequacy. A change in dimension may take place when different concepts are paired with different scales. Use of the Measure of Sampling Adequacy can assist the researcher in identifying which scales are inferior, and which scales to retain. Dickson and Albaum (1977) tested the concept x scale interaction of the semantic space by developing bipolar scales where the majority of anchors were phrases. Their scales were found to be reliable at the .01 level of significance. Benel and Benel (1976) investigated whether there were male/female differences in rating the semantic differential for the three-dimensional concept scale interaction (evaluation, activity, potency), and found no differences in rating.

The three-dimensional concept scale interaction for the semantic space appears to be robust across studies, although it cannot be counted on to hold for ideosyncratic differences. Different scales combined with different concepts may not prove adequate. For different experimental conditions, the researcher is forced to design new instruments instead of borrowing instruments from different investigators (Vidali, 1976; Dickson & Albaum, 1977; Dziuban & Shirkey, 1980).

The literature reviewed indicated the subjects were all students with the exception of Dziuban and Shirkey (1980), where the subjects were school teachers. The field of marketing research has used semantic differential scales to design questionnaire surveys more than any other type of scale (Dickson & Albaum, 1977; Downs, 1978). Prior to application of these scales to the Armed Services, research using semantic differential scales would require scale developmental procedures using the military population to construct scales specific to their research situations.

Since the largest portion of the variance has been found to be in the underlying structure, termed evaluation component of the semantic space, there is always the possibility of a socially desirable response set. This tendency is especially pronounced for the application of the semantic differential to measure personality traits where trait and desirability dimensions become confounded. Klockars (1979) determined that subjects had a stronger tendency to select adjectives which were opposite in desirability when a socially undesirable adjective anchor was presented first. This finding was significant at the .05 level. Klockars, King, and King (1981) were not able to substantiate a social desirability response set where bipolar scales were anchored by adjectives. The anchors were not correlated with social desirability. The inclusion of, or lack of, a social desirability response set may be associated with the developmental procedures used in selection and configuration of the semantic differential scales.

Typically, semantic differential scales include seven scale points with anchors at each end, although Albaum, Best, and Hawkins (1981) and Vidali (1976) developed scales with five scale points. Albaum, Best, and Hawkins (1981) reported a review of the literature where McKelvie (1978) indicated a loss of information when scales employ fewer than five or six scale categories. No further gain of information was obtained beyond 9 to 12 categories. These findings are consistent with other literature that indicates modification of format can produce similar results among instruments.

In research performed by Dolch (1980), two semantic differential scales were developed. One was a numerical scale and the other was an adverb scale. The correlation between the two scales was +.929. It appeared that the format difference made little difference in the response distributions. Yet, a factor analysis revealed that the meaning in the semantic space was not equivalent for the two instruments.

In general, the semantic differential scales have consistently maintained acceptable levels of reliability among studies. Validity has not always been measured. This finding is based on the premise that sound developmental procedures are used in scale construction. Semantic differential scales have appropriately been used in different contexts to measure the meaning of words and attitudes. The scale is flexible in measuring different concepts, and can be applied successfully in a number of environments.

Conclusions Regarding Semantic Differential Scales

As with other scales, no one semantic differential format has proved superior to others. Even though three primary concepts constitute the semantic space, the true meaning of the semantic space may not be known (Dolch, 1980). Of course, it is possible that there never will be a true meaning for the semantic space since semantic meanings change over time. In addition, semantic meaning is dependent on the spoken word and the written word, which are both interpreted by the encoding and decoding of the subject. It also follows that any addition, deletion, or other type of modification would have the potential to change the meaning of the semantic space.

The issue of social desirability response sets may be overcome by careful scale construction (Klockars, King, & King, 1981; Klockars, 1979). The use of the semantic differential scale has received extensive research. Support for this type of scale has been indicated by research results that consistently produced levels of significance at the .05 level and above (Albaum, Best, & Hawkins, 1981; Malhotra, 1981; Mann, Phillips, & Thompson, 1979; Downs, 1978; Dickson & Albaum, 1977; Vidali, 1976).

Downs (1978) administered three versions of the semantic differential. Versions A and B, the nontraditional semantic differential scales, are illustrated in this section under Examples of Semantic Differential Scales. While finding no difference among the response distributions to the three versions, the traditional version was preferred by the respondents. The semantic differential is sensitive enough to measure person, product, and self-concepts so that it can be used to coordinate the image of a product to a target market (Malhotra, 1981). The semantic differential scale can be used in many environments, is flexible as to alterations in the format, and holds fairly stable to the three-dimensional semantic space. However, these studies on the semantic differential do not reflect the operational test and evaluation community's concern for the evaluation of weapons systems. It may be feasible to research the application of the semantic differential scale to this type of environment. Respondent attitudes toward equipment would be a viable area of application.

2.4 RANK ORDER SCALES

Description of Rank Order Scales

Rank order scales originate from ordinal scale measurement. The categories on a rank order scale do not indicate how much distance there is between each category, and unequal distances are assumed. The ranking process by the respondent establishes a hierarchical order (Orlich, 1978), which is also an ordinal order. In the development of rank order scales for survey use, subject ranking has been commonly used (Backstrom & Hurchur-Cesar, 1981). Respondents receive instructions on the assignment of numbers to the items (1, 2, 3, 4, etc.). This is to reveal the rank ordering that the respondent places upon the item in terms of an attribute, such as beauty, length, performance, and preference. It is possible that there may be any number of dimensions along which the respondent is asked to rank order things. This set of rank orderings is termed the ordinal set so that a rank order scale is synonymous with an ordinal scale.

Thurstone investigated rank order scales and how to compare psychological variables. He developed the law of comparative judgement with an underlying assumption which is defined in the following way: "... the degree to which any two stimuli can be discriminated is a direct function of the difference in their status as regards the attribute in question" (McIver & Carmines, 1981). Thurstone generated three new scaling methods based on his law of comparative judgement. The three scaling methods are known as paired-comparisons, successive intervals, and equal appearing intervals.

Rank order scales continue to be used in survey research, although other scaling methods have gained popularity, such as Likert and Guttman scales. There have been instances when rank order scaling procedures have been integrated with other complex systems. An illustration of this is the delta scalar method used by the U.S. Navy and the Air Force Aerospace Medical Research Laboratory. The delta scalar method is a complex system of rank ordering found in the Mission Operability Assessment Technique and Systems Operability Measurement Algorithm (U.S. Navy), and the Subjective Workload Assessment Technique (U.S. Air Force) (Church, 1983). These systems involve establishing a rank order scale that is converted to an interval scale (converting ranked data into an interval scale is sometimes incorporated into the developmental procedures for behaviorally anchored rating scales (BARS) and behavioral expectation scales (BES).

Shannon and Carter (1981) combined rank order methods with 7-point and 5-point scales to measure pilot training. Shannon (1981b) designed a battery to assess aviator performance for pilot training on propeller, jet, and helicopter aircraft. A behavioral analysis was performed using task analysis that included procedures such as rank ordering to isolate the critical components of the task. In other research performed by Shannon (1981c), questionnaires were mailed to all operational squadrons in the fleet using two 7-point functional inventory scales to measure: time, effort, importance of each task, duty, and role. After the data from the questionnaires was quantified, the tasks were rank ordered. It was felt

that this type of procedure would enable the researchers to identify specific tasks which required addition, deletion, or modification for training purposes.

Rank ordering is, therefore, used in questionnaire research in two ways: by developing rank order scales which stand alone, or by embedding rank ordering into the developmental procedures of more complex scales.

Examples of Rank Order Scales

An example of a rank ordered questionnaire item used in computer based instruction research is provided. In this example, the respondent is to rank each statement by descending order of preference.

> What aspects of computer aided instruction did you especially like? Please rank order the following statements using each choice only once.
>
> ___ Courseware is well designed for instructional purposes.
>
> ___ Diagnostic testing and prescriptions meet course objectives.
>
> ___ Student progress reporting is used as an integral part of the training program.
>
> ___ Proctor assistance provides savings in the amount of time required for training.
>
> ___ Students progress at an individual pace to resolve technical problems assigned to them.

Rigney, Towne, Moran, and Mishler (1980) use a ranking by preference for number of hours to practice on system troubleshooting (on a Generalized Maintenance Trainer-Simulator and on actual equipment).

> "If I had 10 hours to practice system troubleshooting, I would divide my time as follows between GMTS and the actual SPA-66 Radar Repeater:"
>
> ___ "hours on GMTS"
>
> ___ "hours on actual equipment"
>
> Total = 10

Comparisons of Rank Order Scales

Rank order items are used in questionnaires that deal with a variety of applications, such as: marketing research (Reynolds & Jolly, 1980), educational research (Orlich, 1978), public opinion polls (McIver & Carmines, 1981), and military research (Church, 1983).

Reynolds and Jolly (1980) compared three different scale methods for reliability (rank order, paired-comparison, and a rating scale with a

Likert format). Analysis of the data for test-retest reliabilities varied depending on whether a Spearman rho was used or Kendall's tau. Using Spearman's rho, the three methods appear to have equal reliabilities. They recommend the use of Kendall's tau as a more appropriate measure of reliability. Using Kendall's tau, the rank order and paired-comparison procedures are more reliable than the rating scale method. They found that the rating scale and rank order technique required less respondent time to rate than paired-comparison (significant at the .0001 level). Their findings would indicate that rank ordering would be a preferred scale format.

Most questionnaire items today are based on formats other than rank ordering (e.g., Likert scales). There is not enough research evidence to substantiate the use of rank order scales in place of other scaling methods.

Conclusions Regarding Rank Order Scales

Rank order scales are appropriate for survey items dealing with ordinal measurement. When Thurstone developed the law of comparative judgment, his scaling techniques were considered a major advancement. Since rank order scales and paired-comparison scales both have a foundation in ordinal measurement, rank order scales would be more time and cost effective than paired-comparison scales.

Current research indicates that the use of rank ordering is in transformation because it is being used and embedded in the procedures of more complex scaling systems (Church, 1983; Shannon, 1981b, 1981c). More research will be required to determine how functional, reliable, and valid these new procedures will be. For example, the statistical analyses achieve varying results when the ordinal data is converted into interval scales. Some of the new scaling systems require prolonged periods for scale development (Church, 1983).

2.5 PAIRED-COMPARISON ITEMS

Description of Paired-Comparison Items

The development of scales, using the paired-comparison method, has been applied to many situations, such as: performance appraisal, opinion surveys, marketing research, food technology, and sports competition (Edwards, 1981; McIvers & Carmines, 1981; Bradley, 1982).

Paired-comparison methods were developed by Thurstone. He proposed systematic procedures for attitude measurement based on the law of comparative judgments. The Thurstone law of comparative judgments includes three different procedures for scale development which include paired-comparisons, successive intervals, and equal appearing intervals. The underlying assumption for the law of comparative judgments is that for each variable measured, there is a most frequently occuring response (McIvers & Carmines, 1981).

In the application of the method, respondents are required to compare several alternatives. Each item is compared with every other item, and results in an overall ranking. Comparison of more than 10 items would be disfunctional since it would require more than 45 separate combinations taken two at a time (Backstrom & Hurchur-Cesar, 1981).

Examples of Paired-Comparison Items

This survey item is constructed to compare an individual's preference for executive performance characteristics. The respondents have three response alternatives to compare: (1) versus (2), (2) versus (3), and (3) versus (1).

For superior executive performance, which behaviors do you find to be most needed?

(1)	Has many meetings and discussions with associates	or	(2)	Usually decides and takes action quickly
(2)	Usually decides and takes action quickly	or	(3)	Usually follows suggestions made by subordinates
(3)	Usually follows suggestions made by subordinates	or	(1)	Has many meetings and discussions with associates

Edwards (1981) developed a modification of the paired-comparison item in which he presented multiple pairs of comparands (people to be rated) at the same time. He enlarged the rating alternatives available to the raters from three to five. He used his new format in an effort to improve performance appraisal. His system of appraisal uses raters to make comparison

ratings about the performance or potential of two individuals on one criterion at a time, and preserves the previous ratings for the judge to consider as he rates additional pairs of ratees. Edwards felt that ratees accepted this approach to performance appraisal since it was more credible to compare peers on a job than to compare an individual against an abstract or vague standard. Edwards uses the following example of three ratees on the criterion "ability to develop people."

	Much Better	Somewhat Better	About Equal	Somewhat Better	Much Better	
Ruth Sproul	( )	( )	( )	( )	( )	Dan Parker
Dan Parker	( )	( )	( )	( )	( )	Ron Half
Ron Half	( )	( )	( )	( )	( )	Ruth Sproul

Comparisons of Paired-Comparison and Other Items

Reynolds and Jolly (1980) and Landy and Barnes (1979) compared a graphic rating scale, a Likert scale, and a rank order item to a paired-comparison item. Each study used college students as subjects. Scale anchors were Behaviorally Anchored Rating Scales (BARS) and Likert anchors, and were from 1, "not at all important," to 7, "extremely important." There were seven scale points for Likert and BARS formats. The results from both studies indicated that different scaling techniques produce different results. It has not been determined which scaling technique is more accurate.

Reynolds and Jolly (1980) reported the work of Munson and McIntyre (1979) where several findings were made about the reliability of assigning numerical ranks, Likert ratings, and an anchoring approach used by respondents. In the anchoring tasks, respondents had to position values at the 1 and 7 points on a Likert scale. Munson and McIntyre found that the anchoring approach was significantly lower in test-retest reliability than assigning numerical ranks. They also found that the Likert scale was lower in test-retest reliability than the ranking procedure, but not significantly less reliable. Munson and McIntyre suggested replacing the rank ordering procedure (originally recommended by Rokeach, 1973) with a Likert rating scale. A reversal of this finding was discovered when Reynolds and Jolly subjected their data (value profiles used in market segmentation) to Kendall's tau instead of Spearman's rho. They found the graphic rating scale method to be significantly less reliable than paired-comparison or rank ordering.

In the development of BARS, other rating procedures serve as preliminary techniques before the final product is constructed. Landy and Barnes (1979) used a graphic rating procedure and compared it with a paired-comparison procedure to assist them in identifying items that would be used later in two different versions for BARS. The BARS development procedure requires that individuals make absolute judgments about the desirability of potential anchors for their place along the scale line (see Section 3.1, Behaviorally Anchored Rating Scales). It has been suggested that the BARS development procedure would be improved by having individuals use comparative judgments instead of the usual absolute judgments about anchors. They discovered that these two different procedures produced different results. The paired-comparison procedure produced end anchors with higher

variances, and middle anchors with lower variances. There were more data points per anchor with the paired-comparison procedure. It appeared that the paired-comparison dispersion of anchors along the BARS scale line produced better estimates of the population than the graphic estimates identified by the other procedure. Using the paired-comparison procedure is a possible way to generate more anchors for the center of BARS.

The literature reviewed on paired-comparison items seems to indicate that different scale development methods result in different item variances. However, the correct scale values are not known when comparing different types of items and scales. In addition, the results are mixed as to which is the most reliable item and scale, depending on which type of statistical analysis is used. According to McIver and Carmines (1981), it has not been possible to provide evidence of unidimensionality for the Thurstone scaling method.

Conclusions Regarding Paired-Comparison Items

Based on the current research, it is not possible to substantiate the use of paired-comparison items as being superior to other types of items and scaling methods.

One of the drawbacks to using this kind of item is that, when more than 10 items are compared, it can become confusing to the respondents (Backstrom & Hurchur-Cesar, 1981). It is also time consuming to use paired-comparison items. Reynolds and Jolly (1980) found that rank order scales and Likert scales required less respondent time to complete than did paired-comparison items. This difference was significant at the .0001 level.

Some researchers promote the use of a rank ordering type of scale (Edwards, 1981; Bradley, 1982). Rank ordering has been suggested for use in performance appraisal and market research. However, rank order scales have fallen out of usage with most types of survey research. An illustration of this lack of usage is that public opinion surveyors more or less abandoned paired-comparison items in the construction of surveys to measure the political system. The paired-comparison method was quite popular in the 1920s and 1930s (McIver & Carmines, 1981).

2.6 CONTINUOUS AND CIRCULAR SCALES

Description of Continuous and Circular Scales

Researchers have examined the equivalence of information obtained by various scale formats. In the search for reliable and valid scales, continuous scales that have no scale points have been compared to more traditional scale formats, such as: semantic differential scales (Albaum, Best, & Hawkins, 1981), rating scales with different numbers of categories (5 through 11) (McKelvie, 1978; Oborne, 1976), and different types of anchors (phrases, adverbs, and color shading) (Oborne, 1976; McKelvie, 1978; Lampert, 1979).

The rationale for comparing continuous scales to other formats has been that a continuous scale will yield greater discrimination by raters. The application of continuous scales has been wide and varied. As an illustration, continuous scales have been used in ergonomics to rate perception of a thermal stimulus (Oborne, 1976). Continuous scales have been used in an opinion survey for satisfaction with respondent's job and apartment (Lampert, 1979). In the latter research, the continuous scale consisted of a rectangular opening within a housing that contained a moving colored bar. White and black represented the two extremes of the scale.

McCormick and Kavanagh (1981) scaled items on an Interpersonal Checklist to a circular scale model. Originally, Guttman (1954) proposed that psychological tests and scales could be related to each other in a circular structure which was termed circumplex. The procedure for scaling in a circular structure may have advantages over paired-comparisons and multidimensional scaling since more stimuli can be scaled. Errors of extreme judgments and central tendency may be eliminated (McCormick and Kavanagh, 1981).

Disadvantages associated with transforming items to a circular model have tended to be the displacement of items from their original organization due to the circular scaling procedures. This phenomenon appeared to be caused by differences in the intensity of the dimension where items were pulled away from the dimensions they originally were intended to represent. Analysis at the item level indicated that items tended to cluster into new factors. McCormick and Kavanagh (1981) suggest that this may be a favorable outcome since the circular scaling procedures can be used to study item ambiguity and item discrimination. These different scaling procedures (circular and bipolar) provide different interpretations for the meaning of items.

Examples of Continuous and Circular Scales

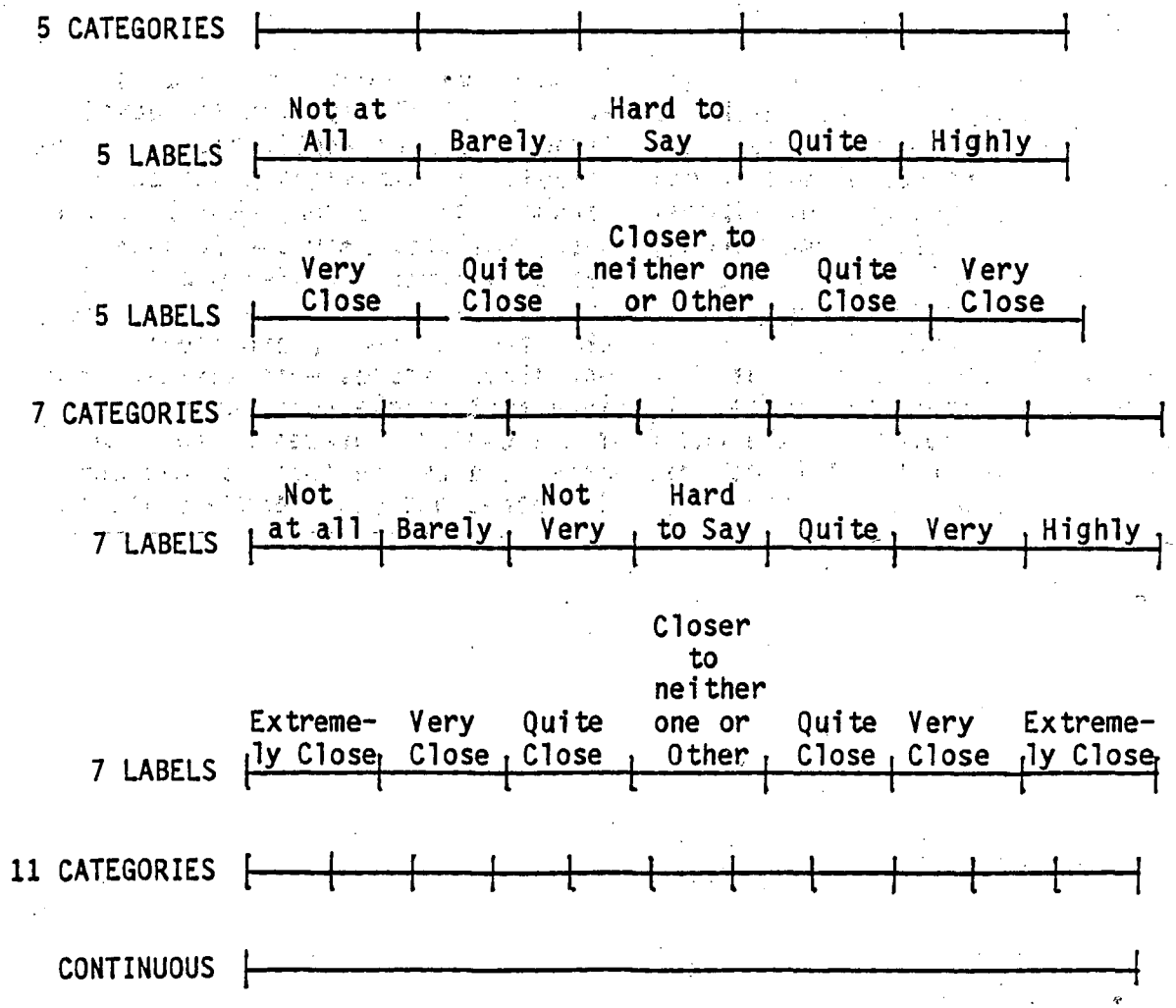
Continuous scales are usually thought of as straight lines with no indications of any differentiation along the scale lines. A continuous scale can provide the respondent with guidance as to the directionality of the rating, and offer the respondent greater discrimination as to ratings along the scale line.

Albaum, Best, and Hawkins (1981) examined the equivalence of data obtained from a continuous rating scale and a semantic differential with five scale points. The distance between the polar opposite terms was 125 mm for both formats. In order to compare the two scale formats, they used university students as subjects to assess their University, Student Union, and University Bookstore. Following is an example of the continuous and discrete scales from Albaum et al.:

Friendly . . . . . Unfriendly

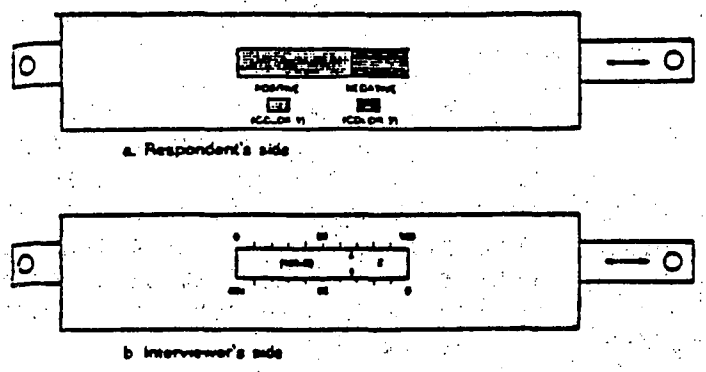
Friendly _ _ _ _ _ _ _ Unfriendly

In research performed by McKelvie (1978), continuous scales were compared to discrete scales with 5, 7, and 11 scale points. The continuous scale consisted of a 16.5 cm line, and the discrete scales were of approximately the same length. Subjects used the scales to make two types of judgments. They were to assess which of 10 adjectives was most descriptive of French Canadians, in general, relative to English Canadians. Ratings to the left of the midpoint meant that the adjectives were less descriptive of French Canadians. Subjects were also asked to take a tone test where they had to rate the pitch of 10 pure tones. An illustration of McKelvie's scales is provided for scales used to measure tone and perceptions of French Canadians/English Canadians.



A new twist on the continuous scale format was developed by Lampert (1979) where a housing with a rectangular opening exposes a color bar that moves in the housing. Lampert termed this device the Attitude Pollimeter. Any topic can be rated by moving the color bar between two colors. One color represents the positive, and one color represents the negative. Subjects using the Attitude Pollimeter answered 10 questions related to satisfaction with their apartment and job. A diagram of Lampert's Attitude Pollimeter is presented here:

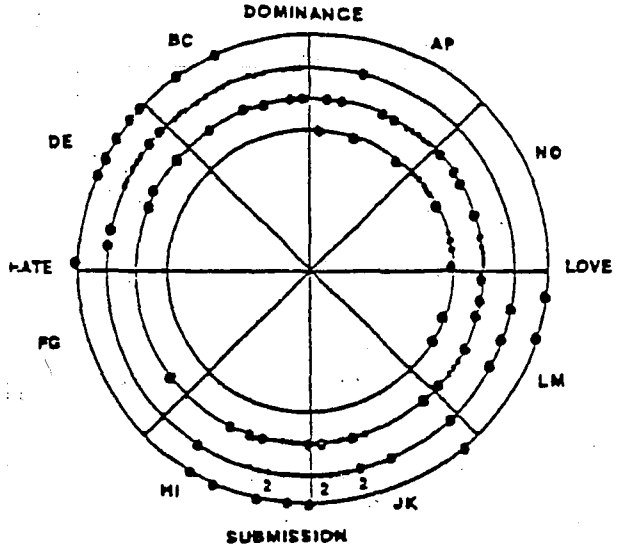
THE POLLIMETER (PATENT PENDING)



The circular scale has been found in many assessment areas and is known as a circumplex. McCormick and Kavanagh (1981) reported the development of empirical circumplices for a large number of applications. A few of the examples are as follows: Wecksler-Bellvue Intelligence Scales (Guttman, 1957), Minnesota Multiphasic Personality Inventory (MMPI) (Schaefer, 1961; Slater, 1962), and Strong, Kuder, Holland, American College Testing Program (ACT) (Cole, 1973). Based on Guttman's (1954) model, McCormick and Kavanagh scaled personality items into a circular structure. In the generation of a circular scale, 128 items on the Interpersonal Checklist (ICL) were rated. The four concentric circles were divided into eight equal pie-shaped intervals. The innermost circle represents mild items (ICL). The second circle out from the center represents moderate items (ICL). Strong (ICL) items are represented by the third circle out from the center, and the outermost circle represents (ICL) extreme items.

Following is an example of McCormick and Kavanagh's (1981) circular scale for the Interpersonal Checklist in the dimensions dominance-submission and love-hate:

Angular Item Placement of the 128 Items
of the ICL from the Two-Dimensional Scaling Procedures



As can be seen, researchers have multiple options for scale layout. Continuous and circular scales were presented as an illustration to expand researchers' options beyond the traditional bipolar scale. The circumplex avoids the problem of errors of central tendency.

Comparisons of Continuous and Circular Scales

In the comparison of these scales, most of the research conducted has been with college students as subjects, with the exception of eligible raters in the City of Jerusalem (Lampert, 1979) and British Rail passengers using intercity trains (Oborne, 1976).

It is not possible to make a clear comparison of the continuous scales since each investigator's concept of a continuous scale is different. In addition, the comparison of continuous scales to other types of scales varied with each study. As an example, Albaum, Best, and Hawkins (1981) compared a 125 mm continuous scale to a semantic differential scale. Both scales were anchored by adjectives. McKelvie (1978) compared a continuous scale to scales with 5, 7, and 11 scale points. McKelvie's scales were all approximately 16.5 cm in length. Tones and opinions were both measured.

The results of the research on continuous scales seems to indicate that it is possible to develop and apply a continuous scale without affecting the psychometric properties of the scale. Continuous scales appear to be equivalent to traditional scales with discrete categories. Albaum, Best, and Hawkins (1981) achieved r = .95 between continuous and semantic differential scales. McKelvie (1978) found that reliability was unaffected by scale type when continuous scales were compared to category scales.

Subjects using continuous scales appeared to be effectively using what would be equivalent to five categories on the adjective task and six categories on the tone task. There was no evidence that the continuous scales were more reliable or valid than the category scales, although subjects stated that they preferred the continuous scales. Subjects perceived that they performed more consistently and accurately with the continuous scale.

Of particular interest is the research performed by Oborne (1976). The focus was on the development of rating scales applied to field studies in ergonomics. Oborne combined two scale development procedures for continuous and category scales. This combination came about because the investigator felt that ratings along a continuous scale could not be accurately transformed to a numerical equivalent, and that category scales were ordinal measures. Oborne transformed the continuous scale from a psychophysical measuring instrument into the beginning of what was termed a "comfort indicator." This procedure was accomplished by analyzing the spread of ratings along the continuous line, and then reducing the data into five groups of categories. Descriptive phrases were then developed for each category in the second phase of scale development. Ratings were obtained for noise intensity, vibration intensity, and comfort. Obsorne's (1976) unique approach to combining developmental scale procedures to include both continuous and category scales may be useful in the measurement of psychophysical phenomena.

In a comparison of four scales, the Attitude Pollimeter (Lampert, 1979) (a continuous scale), bipolar continuous scale, numerical scale, and verbal scale, the means and standard deviations were similar for three of the scales. Apparently, the scale format had little effect on the statistical measures for the two continuous scales and the numerical scale. The verbal scale was an exception. The correlation coefficients between the Attitude Pollimeter and the numerical scale were highest at $r = .929$, and lowest for the verbal scale at $r = .888$. The differences were significant at the .001 level among all correlation coefficients. Three of the instruments were based on a 0 to 100 scale, while one of the instruments (the verbal scale) was on a 1 to 5 scale. The verbal scale had five categories from "very satisfied" to "very unsatisfied." The bipolar continuous scale consisted of a continuous line which was anchored at each end by "very satisfied" and "very unsatisfied." The numerical scale ranged from 0 to 10. The actual recording and conversion of responses was as follows: the Attitude Pollimeter recorded responses from 0 to 100; the numerical scale converted responses 0 to 10 into 0 to 100; and the bipolar continuous scale converted responses along the continuous line from 0 to 100. He determined that the measurement procedure had little effect on the statistical results. The variances for the continuous scales (bipolar and Attitude Pollimeter) and discrete scale (numerical) were about the same. This suggests that respondents were continuing to avoid the use of extreme ratings even when using a continuous scale. The verbal scale was rated by a plurality of the respondents (40.9%) as being best, easiest, and most pleasant to use out of the four scales.

Continuous scales appear to be psychometrically as sound as the more traditional scale formats, but it has not been possible to establish their superiority over other scale formats.

Conclusions Regarding Continuous and Circular Scales

Continuous scales offer the researcher another option in the selection of scale format. Even though some researchers prefer the use of continuous scales to offer the respondent a greater differentiation in rating, this may not necessarily be realized. For example, Oborne (1976) found respondents rating continuous scales on an equivalent of five categories. McKelvie (1978) found respondents rating continuous scales on the equivalent of five or six categories depending on what was being measured.

Even though subjects may state that they prefer a continuous scale over a category scale (McKelvie, 1978), their preference does not indicate pychometric superiority of this format over other formats. In a comparison of four scaling formats, Lampert (1979) found that subjects with a low level of education preferred verbal scales as their first choice, with the Attitude Pollimeter (a continuous scale) as their second choice.

Since format variations do not seem to influence the psychometric results to that great of a degree, there are some novel developmental scaling procedures open to researchers. One of these is in the area of ergonomic measurement. Further research in ergonomic scale development seems reasonable for the integration and transformation of continuous scales into category scales in the measurement of psychophysical phenomena using the procedures of Oborne (1976).

For respondents who have a low educational level, the Attitude Pollimeter may be appropriate. There may be a drawback to its use in large surveys (Lampert, 1979). In using the Attitude Pollimeter, each respondent is interviewed. Rating takes place on a one-to-one basis with this device. At its present level of development, this would not be a cost effective approach to obtaining survey data.

One of the most unusual approaches to scale development has been the circular scale (also known as the circumplex). Circular scales have been developed for almost every area of psychological assessment, such as: intelligence tests, personality inventories, and vocational inventories (McCormick and Kavanagh, 1981). One of the advantages for using circular scales has been the elimination of the error of central tendency through random presentation of the stimuli. Another advantage is the measure of variability of response to an item which allows for the determination of item ambiguity and discrimination. One of the drawbacks to this scale is the skewness in item distributions brought about by the procedure.

Continuous and circular scales appear to be as effective as other scales. Considerable effort is required for the development of circular scales. The selection of scale format is essentially based on the preference of the investigator. As with other types of scales, the developmental procedures have greater importance than the scale format.

CHAPTER III

BEHAVIORAL SCALES

Behavioral scales are reviewed in this chapter for Behaviorally Anchored Rating Scales (BARS), Behavioral Expectation Scales (BES), Behavioral Observation Scales (BOS), and Mixed Standard Scales (MSS). There are a wide variety of behavioral scales using variations of the Smith and Kendall (1963) format. This list of behavioral scales is not claimed to be all inclusive. Behavioral scales were developed to encourage raters to observe behavior more accurately. The primary application for these scales has been in the area of performance appraisal. Other applications have emerged since they were originally established by Smith and Kendall. Behavioral scales are built on critical incidents, and they have been used to: evaluate morale, establish feedback, train raters, and delineate organizational goals. They could be used as a link to Management By Objectives (MBO) during the planning stage.

The time and cost factors involved in developing behavioral scales has been extensive compared to other scaling techniques. To make this scaling technique viable, it may be necessary to generalize the use of the behavioral scale to multiple applications such as those mentioned above. In addition to this constraint, psychometric studies of behavioral scales have not indicated that they are consistently better than other types of scales.

Psychometric soundness for these scales has depended largely on the specific developmental procedures used. For example, critical incidents are grouped into dimension categories by groups of participants. The percentage level of agreement for inclusion of a critical incident into a dimension varies with different research projects. It may fluctuate between 60% up to 80% agreement depending on the research method. To improve accuracy in ratings, training sessions have been used so that raters would better understand how to use behavioral scales and how to evaluate performance. Training raters to reduce errors has brought about mixed results. The amount of time devoted to the training, as well as the content of the training, have influenced ratings using behavioral scales. An illustration of the varied impact of training was reported by Bernardin and Walter (1977) where halo error was reduced, but ratee discrimination and inter-rater reliability were not increased by training.

Some of the varied approaches to developing behavioral scales have had their own inherent problems. BES translate actual behaviors into expected behaviors. This procedure culminates in requiring raters to infer a ratee's ability by predicting what the ratee's expected performance will be. Another deficit has been in the content of BOS. BOS use critical incidents to define effective and ineffective behaviors. There is the possibility that some of the behaviors may be exhibited so infrequently that they are not useful in differentiating among ratees. MSS were developed to reduce rating errors by randomizing the presentation of items. This has been frustrating to some raters. MSS has an apparent lack of face validity, yet at the same time is internally consistent.

Behavioral scales are designed to rate the performance rather than traits of individuals. It is possible that judgments made using these scales require recall of performance over extended periods of time. This indicates that behavioral scales may be measuring traits as well as behaviors.

3.1 BEHAVIORALLY ANCHORED RATING SCALES

Description of Behaviorally Anchored Rating Scales

A wide variety of forms and methods of scale development is grouped under the term Behaviorally Anchored Rating Scales (BARS). BARS were established to encourage raters to observe behavior more accurately (Bernardin & Smith, 1981). These scales have developmental procedures based on the Smith and Kendall (1963) format. The original developmental procedure established by Smith and Kendall had six steps. Subsequent researchers have slightly varied the original developmental methodology with successive refinements (Murphy, 1980).

It was recommended by Smith and Kendall (1963) that the rating environment remain constant across ratings, and that the raters rate the ratees in a similar manner (Jacobs, Kafry, & Zedeck, 1980). The raters are required to make inferences from observed behaviors to expected behaviors. This allows the rater to generalize from the specific critical incidents listed out on the BARS form to the range of equivalent incidents that the rater has observed regarding the behavior of the ratee while they work together on the job. The expected behaviors are those that are printed on the BARS format. In addition, it is not possible to list out every possible expected behavior on a BARS format. The rater must generalize to what would be expected behavior by the ratee. This is based on the transformed critical incidents identified along the scale line on the BARS format. It is assumed that the rater will be able to review the behavioral anchors and select the behavioral anchor which best represents the expected behavior of the ratee.

A potential rating problem may occur if the rater has not observed the behaviors identified by the behavioral anchors. The task of the rater is to generalize from known behavior to what would be an expected behavior when the anchors do not adequately describe the ratee. Rating expected behaviors may facilitate rating unobserved events (Jacobs, Kafry, & Zedeck, 1980). Behavior-based scales appear more reliable than trait- based scales for performance appraisal measurement. BARS are more specific in identifying behaviors (Schneier & Beatty, 1979a, 1979b, 1979c) observed on a job than a personality trait such as "responsibility."

As a way to minimize error in scale development, Bernardin, La Shells, Smith, and Alvares (1976) suggested that each dimension of performance be defined with critical incidents for each interval on the dimension. They used two groups of participants for selecting and scaling the critical incidents. The first group placed critical incidents into dimension categories with at least 60% agreement for incidents to be included. Researchers initiate this process with hundreds of critical incidents, and at times there may be over 1,000 critical incidents. Individuals are selected as judges, and they are required to make a judgment as to which dimension a critical incident would fit into. If the 60% level of agreement has not been reached, then the critical incident would be deleted from the pool of

critical incidents. This is a way to reduce the number of critical incidents used in the BARS format. The second group of judges rates the critical incidents regarding the value of behavioral dimensions. It is desired that critical incidents which form a dimension have as little overlap as possible with the other dimensions (Landy & Barnes, 1979). The first group of participants working with the critical incidents will typically generate enough incidents to establish between 5 and 12 dimensions (Cocanougher & Ivancevich, 1978).

Anchoring the critical incidents to the scale continuum may affect the means and standard deviations as different scaling procedures are used. Locander and Staples (1978) and Staples and Locander (1975) advised anchoring the critical incidents to a 10-point scale with 9 intervals. The values assigned to the scale were between 0.00, "undesirable," to 9.00, "highly desirable." The degree of effective or ineffective job performance was assigned to behaviors which were subsequently scaled between 0.00 and 9.00 for each performance dimension. Each behavior (critical incident) along the scale line was analyzed for mean scores and standard deviations. Paired-comparison and graphic rating scale techniques have been used to anchor critical incidents to intervals on dimensions (Landy & Barnes, 1979).

BARS has typically been used to construct scales for performance appraisal. In an effort to construct a scale to measure morale in military units by means other than self-reports, Motowidlo and Borman (1977) were able to successfully use BARS. They developed eight dimensions of group morale to rate 47 platoon-sized units in the U.S. Army stationed in a foreign location. Even though the critical incidents are from different jobs within the Army, they reflect the morale of the soldiers. The Motowidlo and Borman BARS format for the dimension "performance and effort on the job" covered the morale level for a variety of jobs. This means that the rater must generalize to the expected behavior since not every description of morale for each type of job is part of this scale. This is an illustration of the Jacobs, Kafry, and Zedeck (1980) warning about generalizing from a critical incident to an expected behavior (see Section 2.1, Multiple-Choice Scales).

Examples of Behaviorally Anchored Rating Scales

An example below is from Motowidlo and Borman (1977). BARS have traditionally been developed for performance appraisals. In an unusual application of BARS, Motowidlo and Borman developed a scale to measure morale for military units stationed in the U.S. and in two foreign locations. The strategy used was to obtain examples of expressions of morale. They started out with 1,163 examples of morale. The BARS illustrated here represents behavioral anchors associated with "performance and effort on the job." Each scale point is designed to reflect a different level of morale. A high level of morale indicates behaviors such as spending extra time to get the job completed and volunteering to perform the task well.

Scale Dimension

Performance and Effort on the Job

Scale Point	Behavioral Anchor
9	When maintenance mechanics found an error in their assembly procedures on an aircraft, they told their platoon leaders of their mistake and requested that the hangar be open Saturday and Sunday if necessary to meet their previously promised Monday delivery.
8	While clearing the brush from an approach to an airport, these dozer operators never shut the dozer off, running in shifts right through lunch.
7	This section was asked to prepare a set of firing charts by a specific time. The charts were finished ahead of time.
6	Although this section was constantly called upon for typing tasks, the work was done with few mistakes and on a timely basis.
5	The men in this unit did not push for top performance, although they did their jobs and kept busy.
4	Many troops in this unit would leave the post as quickly as possible after duty hours to avoid doing any extra work.
3	The service section of a support unit had a large backlog of equipment needing repair. All enlisted personnel assigned to this section appeared to be busy, but their output was very low compared to the other service sections.
2	The men in this section signed out weapons to be cleaned but sat around and "shot the bull" until it was time to turn the weapons back in.
1	During one period these enlisted personnel slowed their work down and made mistakes that cost time and new parts. They were working 7-day weeks, but at the end of the period they were accomplishing only the same amount of work in 7 days that they had been accomplishing before in 5 days.

Comparisons of Behaviorally Anchored Rating Scales

BARS have been developed for various populations, such as police officers (Landy, Farr, Saal, & Freytag, 1976), soldiers (Motowidlo & Borman, 1977), and students (Hom, De Nisi, Kinicki, & Bannister, 1982). Investigation of BARS has focused on various applications, such as feedback using different instruments as opposed to no feedback (Hom et al., 1982), scaling of critical incidents using paired-comparison and graphic rating (Landy & Barnes, 1979), format differences in conjunction with training or lack of training (Borman, 1979), and the effect of participation in scale construction (Friedman & Cornelius, 1976).

Scale dimensions have ranged from 2 (Landy & Barnes, 1979) through 10 (Hom, De Nisi, Kinicki, & Bannister, 1982). Scale points have numbered between 5 (Hom et al., 1982) and 9 (Landy, Farr, Saal, & Freytag, 1976). Different anchors have been used for critical incidents (Motowidlo & Borman, 1977). For example, numerical anchors along with descriptors "high," "average," and "low" have been used (Landy, Farr, Saal, & Freytag, 1976), as well as non-continuous Likert-type anchors (Hom et al., 1982). Dimensions and anchors have different definitions as well as different titles for the various scale formats.

The Smith and Kendall (1963) model for BARS developmental procedures requires the participation of raters in scale construction procedures. Participation in BARS and graphic rating scale construction has led to increased convergent validity. Participation by raters in scale development did not lead to high levels of discriminant validity (Friedman & Cornelius, 1976). There has been little support for the involvement of raters in scale construction (Kingstrom & Bass, 1981).

Other avenues for the use of BARS have been sought. For example, it is possible to use the data from BARS in the feedback condition for performance appraisal (Hom, De Nisi, Kinicki, & Bannister, 1982). Job analysis can be compared with critical incidents (Atkin & Conlon, 1978). Management by Objectives (MBO) (Locander & Staples, 1978) can be used in conjunction with BARS. Because of the time and money involved in the construction of BARS, the rationale for BARS use without secondary applications provides a weak case for their selection.

The psychometric soundness of BARS has been more promising for developmental procedures than for application in field studies (Jacobs, Kafry, & Zedeck, 1980). There have been disappointing levels of convergent validity, and no discriminant validity for some studies. Mixed results were found for rating characteristics in several types of formats as they were compared to BARS (Borman, 1979; Kingstrom & Bass, 1981).

Many studies have examined the effects of rater training in an effort to reduce rating errors and increase reliability (Landy & Farr, 1980). Using a short training program (5 to 6 minutes), Borman (1975) found little impact on the quality of ratings. Training sessions conducted by Bernardin and Walter (1977) had little impact on ratee discrimination and interrater reliability, but they did reduce halo error. In research performed by Borman (1979), three hours of training versus no training reduced halo error. It did not improve accuracy of ratings. No one scale format was

consistently better than another. Training raters to reduce errors while using BARS has produced varied results.

Errors in ratings may be attributed to a number of sources such as scale format, rater ability to observe behavior, and motivation of raters. Rater effectiveness may also be influenced by the cognitive complexity of raters. Schneier (1977a) viewed BARS as requiring more cognitive complexity than other formats (Jacobs, Kafry, & Zedeck, 1980; Landy & Farr, 1980).

Conclusions Regarding Behaviorally Anchored Rating Scales

BARS psychometric soundness appears to be dependent on the specific developmental procedures used and the research design selected. It has not been possible to substantiate psychometric superiority of BARS. Even so, BARS has not appeared to be inferior to other scales (Murphy, 1980). Specific statistical indices used in different studies created problems of interpretation (Kingstrom & Bass, 1981).

Smith and Kendall (1963) originally recommended the developmental procedures to use in constructing BARS. As more research has been performed in the development of BARS, many investigators have modified the Smith and Kendall methodology to try and improve upon the procedures. It is difficult to compare BARS studies since the developmental procedures vary from study to study. There is the possibility that for some of the procedure, there has been inappropriate matching of rating formats, scales, and raters. This would result in a lack of convergent validity. This is not to say that all such modifications negatively influenced the reliability or validity of the scales.

A serious concern for the development of BARS is the time and cost involved (Cocanougher & Ivancevich, 1978). This is why expanding the use of BARS for more than performance appraisal may be a prerequisite in an effort to capture BARS spin-offs. For example, Staples and Locander (1975) suggest that appraisal criteria may be used as a guide for delineating organizational goals. Another use for BARS could be as a link to MBO during the MBO action planning stage. Performance appraisal dimensions and specific job behaviors can be identified as a way to achieve many objectives (Kearney, 1979). This makes BARS more viable and cost effective to the organization. The application of BARS to measure group morale was encouraging as an alternative to the traditional self-report measures obtained in surveys. This is an indication that BARS is a form of scale construction that can be used in surveys and not only for performance appraisal.

3.2 BEHAVIORAL EXPECTATION SCALES

Description of Behavioral Expection Scales

Behavioral Expectation Scales (BES) were originally derived from the work of Smith and Kendall (1963) for developing behavioral criteria in performance appraisal. BES is based on the critical incident technique where job performance is described. Each observer is requested to provide examples of effective or ineffective behavior. This includes the circumstances that explain what the person did that was effective or ineffective for performance of their job. The critical incidents are grouped into dimensions. If there is not a certain minimum percentage of agreement for assignment to a dimension (usually 60% to 80%), the critical incident is eliminated. Each critical incident is then assigned a scale point which represents: good, average, or poor job performance. The numerical value given to each of the critical incidents is the average numerical rating of all the judges (usually job incumbents participating in scale development are judges). The critical incidents are then used as anchors on the rating scale (Latham, Fay, & Saari, 1979) (see Section 3.1, Behaviorally Anchored Rating Scales).

The resulting scales are known as Behaviorally Anchored Rating Scales (BARS). When the anchors are reworded from actual behaviors to expected behaviors, they are known as BES. Raters are assigned the task of determining whether the behavioral observations of the ratee would lead to the expected behaviors displayed in the anchors along the scale (Latham, Fay, & Saari, 1979).

Examples of Behavioral Expectation Scales

Ivancevich (1980) completed the construction of a BES with a final 6-factor structure using 29 items which represented engineers' attitudes about performance evaluation. Names of the six factors are as follows: "equity (When I am compared to other engineers, my appraisal is fairly determined); accuracy (A major strength of the appraisal program is accuracy); comprehensiveness (The appraisal system covers the total domain of my job); meaningful feedback (I receive information from the appraisal system that helps me determine how I am doing on the job); clarity (The performance dimensions on the appraisal are clear); and motivational (The appraisal system encourages me to correct weaknesses)."

Ivancevich (1980) constructed the scale by attaching seven anchor points to each of the behavioral expectations, for example:

"When I am compared to other engineers,
my appraisal is fairly determined."

Very False 1 2 3 4 5 6 7 Very True

Comparisons of Behavioral Expectation Scales

BES research on scale developmental procedures uses the scaling methodology of Smith and Kendall (1963). Since critical incidents are traditionally assigned to dimensions by percentage of agreement of judges, various researchers have set different percentage cutoffs. Bernardin, La Shells, Smith, and Alvares (1976) manipulated their critical incidents by percentage of agreement for placement in a dimension between 50% and 60% for one scale and 80% or greater for another scale. Some research does not report the percentage of acceptance for inclusion of critical incidents into dimensions (Ivancevich, 1980). Eighty percent appears to be a frequently used criterion (Latham, Fay, & Saari, 1979).

Subjects used in scale development are usually supervisors and subordinates ranging from engineers (Ivancevich, 1979) to semi-skilled workers (Schneier, 1977a), or university students and faculty (Bernardin, La Shells, Smith, and Alvares, 1976; Kafry, Zedeck, & Jacobs, 1976; Bernardin, 1977; Bernardin & Walter, 1977; Fay & Latham, 1982). Borman and Dunnette (1975) expanded this range to include Navy personnel. Their study found that BES reduced rating errors.

Since BES is always tailor-made for a specific organization, the number of dimensions may vary for each scale. The range of dimensions observed was between 14 for Schneier's (1977a) cognitive complex raters, and a more limited number of dimensions, four (Latham, Fay, & Saari, 1979).

The number of scale points varied between and within studies. For example, Beatty, Schneier, and Beatty (1977) compared three scales each having a different numbers of scale points. Their dimensional scale had five points anchored by adjectives ranging from "very poor" to "excellent." A global scale and a BES scale each had nine scale points anchored by adjectives ranging from "excellent" to "unacceptable." Bernardin (1977) compared BES to two summated scales. All three scales had seven scale points. BES and one of the summated scales had behavioral anchors. The other summated scale was anchored by the terms "always" and "never." The number of scale points observed for BES ranged from five to nine.

BES anchors varied according to each study. Kafry, Zedeck, and Jacobs (1976) arranged behavioral anchors randomly into a checklist. After the raters rated the behavioral anchors on the checklist, the behavioral anchors were reconstructed into their original dimensions. The data was subjected to a Guttman analysis to determine whether the behavioral anchors were unidimensional and cumulative. They obtained two different coefficients of reproducibility. The first coefficient was based on the fixed order (the order of anchors originally established by the researchers). The second coefficient, termed the free order, was the best possible order given the responses based on the use of the scales. The Guttman analysis did not indicate a strong unidimensional scale.

The perceptual set of the individuals developing the scale may have been different than the perceptual set of the raters. The raters only observed the anchors in a random order. It is possible this contributed to the lack of unidimensionality and other developmental problems. The judgments about the critical incidents for inclusion or exclusion from the

scale were not made in reference to any one person. However, the raters all used the scale to evaluate a single individual. None of the raters involved in this study participated in the actual scale developmental procedures. Kafry, Bedeck, and Jacobs (1976) suggested that the use of a Guttman scalogram analysis would assist researchers generating BES to identify items and to order the scale. This approach would provide assurity that the scale is unidimensional and cumulative.

Ivancevich (1980) concluded that BES was slightly superior to non-anchored and trait scales in reducing halo error and increasing interrater agreement. In comparing intense training, discussion, and a control group for BES, there were no significant differences in leniency error comparing the discussion group and the control group. Intense training on the BES resulted in significantly less halo error than the discussion group and comparison group (Ivancevich, 1979). Schneier (1977a) found that cognitively complex raters had less halo error than cognitively simple raters for the BES or a simplified alternate version of the BES (see Section 6.2, Cognitive Complexity). Bernardin (1977) compared BES to summated scales and determined that summated scales had less leniency error and greater interrater agreement than BES.

Conclusions Regarding Behavioral Expectation Scales

Nothing conclusive can be said about the psychometric characteristics of BES compared to other rating formats. Researchers have applied many varied approaches to the developmental procedures and formats of BES. Psychometric qualities of BES do not promote its use over more easily developed scales. It appears that BES suffers from judgmental errors and biases. Raters are required to infer the ratee's ability and to predict the ratee's expected performance.

The rigor in developing BES will determine the reliability and validity of the scales more than the format. BES is time-consuming to construct and may not be worth the time or money. There is no clear evidence that BES is superior to other scales unless it can be shown that there are worthwhile by-products, such as clarification of organizational policy, feedback for interviewing in performance appraisal systems, improvement of individual performance, and identification of divergent perceptions of employees.

Thurstone scaling is the foundation for the development of the BES. Thurstone scaling has been used in the past to scale attitudes in the fields of political science and marketing. The construction of Thurstone scales is labor intensive, and judges have difficulty discriminating among the moderate range of items. Public opinion researchers have adapted scaling methods based on Likert and Guttman models. McIver and Carmines (1981) conclude that these models overcome the limitations of Thurstone scaling.

3.3 BEHAVIORAL OBSERVATION SCALES

Description of Behavioral Observation Scales

Behavioral Observation Scales (BOS) use developmental procedures which employ Likert scale methodology. BOS are used to rate the observed relative frequency (or percentage) of occurrence of selected behaviors on a 5-point rating scale. BOS have intervals defined by specified occurrence rates of: 0-65%, 65-74%, 75-84%, 85-94%, and 95-100% (Kane & Bernardin, 1982).

Using a Likert-type rating scale, BOS require raters to identify the frequency with which specific behaviors have been observed over a specified period of time. BOS are built by obtaining a set of critical incidents (Murphy, Martin, & Garcia, 1982). Latham, Fay, and Saari (1979) explain the process as follows: Large numbers of critical incidents are obtained. Individuals are observed and rated for frequency of critical incidents on a 5-point scale. Summing the responses to all the items for each individual provides a total score for each ratee. Item analysis is conducted to identify which items have the highest correlations with the total score on the scale. In research performed by Latham et al. (1979), 514 critical incidents were reported. Critical incidents that were similar in content were collapsed into one behavioral item. This is a frequently used procedure in developing behavioral criteria (Fivars, 1975; Flannagan, 1954). The procedure is repeated many times by correlating items to a criterion.

Examples of Behavioral Observation Scales

Latham, Fay, and Saari (1979) constructed a BOS for first-line foremen and developed a comprehensive description of the foreman's job. They attached a 5-point Likert-type scale to each behavioral item. Foremen were rated by having superintendents indicate on the scale the frequency with which they observed each behavior. An example of a behavioral item for BOS developed by Latham et al. (1979) is provided below.

"Tells crew to inform him immediately of any unsafe condition."

Almost Never 1 2 3 4 5 Almost Always

Comparisons of Behavioral Observation Scales

BOS have been developed for various populations, such as: students (Fay & Latham, 1982; Murphy, Martin, & Garcia, 1982), foremen (Latham, Fay, & Saari, 1979), and logging crews (Latham & Wexley, 1977). The number of subjects ranged from 90 (Fay & Latham, 1982) through 300 (Latham & Wexley, 1977). Researchers have varied the types of experimental conditions by comparing BOS to Behavioral Expectation Scales (BES), trait scales (Fay & Latham, 1982), and graphic rating scales (Murphy, Martin, & Garcia, 1982). The number of dimensions obtained for BOS ranged from 2 (Murphy, Martin, & Garcia, 1982) through 6 (Fay & Latham, 1982). The number of scale points varied between 5 (Latham, Fay, & Saari, 1979) and 7 (Murphy, Martin, & Garcia, 1982). The anchors associated with the scale points changed with

each study, for example, "almost always" to "almost never," and "always," "generally," "sometimes," "seldom," and "never."

Fay and Latham (1982) provided subjects with four hours of training, while Latham, Fay, and Saari (1979) provided subjects with six hours of training. Fay and Latham (1982) found that training led to significantly more accurate ratings than no training. BOS and BES were both significantly more accurate for rating ratees at the .05 level of significance than trait scales were for rating "first impressions" of ratees. The 6-hour training program minimized rating errors for contrast effects, central tendency, positive and negative leniency, halo effect, and first impressions. Latham et al. (1979) determined that BOS was content valid and was capable of differentiating between successful and unsuccessful employees.

In comparing BOS to other scales and conditions, it is not possible to determine or discover any clear trends in the literature. Some of the reasons for this are the lack of replication across studies for: number of subjects, number and types of conditions, number and type of scale points, and number and type of dimensions. Since no one behavioral scale is any less subject to errors than the other scales, the selection of methodology could be based on one's preference for a Thurstone model or for a Likert model, etc. As previously noted, BOS developmental procedures are based on a Likert-type model, and this enhances their psychometric soundness.

Conclusions Regarding Behavioral Observation Scales

Critical incidents used to develop BOS which define effective and ineffective behavior are sometimes observed so infrequently that they lack the ability to differentiate good from bad ratees (Latham, Fay, & Saari, 1979). BOS appear to require raters to make simple observations. This scale may be really measuring a trait like judgment because of the recall over time required by raters (Murphy, Martin, & Garcia, 1982). Another weakness of the BOS is the occurrence rate for each interval. Frequencies for various items of effective or ineffective behavior may not hold constant for each interval with the same percentages (Kane & Bernardin, 1982; Bernardin & Kane, 1980).

Since no one scale format is any less error prone than another, the selection of scale developmental procedures could be based on a preference for the use of a Thurstone scale or a Likert scale. BOS developmental procedures have a Likert foundation which enhances their psychometric soundness. Likert items employ ordinal scales and are primarily used for assessing opinions for survey research. They are also known as summative rating scales and are used to select a set of items that measure the same attitude or attribute (Orlich, 1978). An underlying assumption of Likert scaling is that behaviors of respondents are being rated rather than attitudes. This assumption attributes systematic variation in responses to differences among respondents. Another assumption is that all items, as a group, measure the same attribute so that the sum of the items will contain the same variable as the individual items. Separate scores are treated as predictors of the total scores. However, it has been difficult to substantiate that the sum of the measures collectively measure the same dimension (McIver & Carmines, 1981).

Likert and Guttman scales appear to be superior to Thurstone scales since they have overcome the limitations inherent in Thurstone scales. According to McIver and Carmines (1981), there are three basic assumptions underlying the Likert/summated model: (1) each item has a monotonic trace line, (2) the sum of the trace lines is monotonic and approximately linear, and (3) the set of items measures only the attribute of interest. The use of BOS does not ensure valid ratings. Validity and reliability of scales depends on the rigor of the scale development procedures.

3.4 MIXED STANDARD SCALES

Description of Mixed Standard Scales

Mixed Standard Scales (MSS) are a variant of the Behaviorally Anchored Rating Scale (BARS) technique. MSS ratings are behaviorally based with a high relevance to task performance. It is common to require rater participation in MSS and BARS scale development (Rosinger, Myers, Levy, Loar, Mohrman, & Stock, 1982). MSS were established to reduce rating errors (Saal, 1979). Blanz and Ghiselli (1972) proposed that a reduction in halo and leniency errors would take place by disguising the relationship among the items and the dimensions.

The actual MSS developmental procedures are structured on a Guttman rating method (Saal, 1979). Guttman scaling was developed as a response to deficiencies in scaling techniques established by Thurston and Likert. In a true Guttman scale (McIver & Carmines, 1981), it is possible to predict the subject's response to each item making up the scale. A perfect correlation between overall scale score and item scores is almost never achieved. Guttman scales are able to demonstrate that a series of items belong on a unidimensional continuum. The calculation for scoring the Guttman scale is similar to summing the positive responses on a Likert scale. The divergence between a cumulative Guttman model and a summative Likert model hinges on when the responses are totaled and how the responses are interpreted.

Rosinger, Myers, Levy, Loar, Mohrman, and Stock (1982) described MSS developmental procedures as requiring a 4-step process. Step 1 is a series of interviews with potential respondents for the wording of the three (triad) anchors for each item. The second step consists of taking the preliminary anchor type statements, and having a group of respondents suggest changes for each statement and level of statement in the triad. Feedback by respondents allows for modification of the original statements. The modified statements from the triads are then arranged in a random order. Each of the statements is rated by respondents from 1, "very poor," to 7, "very exceptional." Step 3 requires statistical analysis of the triads to determine which triads to include in the final form. A pilot test of the instrument is performed for Step 4. Since the statements are mixed (and disguised), it is not possible to directly assign numerical ratings to the format.

The respondents must rate each item without knowledge of the item's dimensionality since the items are randomized in their presentation. Each item must be rated with a plus (+), zero (0), or minus (-) (Dickinson & Zellinger, 1980). Items rated with a plus indicate better performance than the item describes. Items rated with a zero indicates that the ratee's performance fits the item description. Items rated with a minus indicate that the ratee's performance is poorer than the item description. When the respondent completes the rating, the ratings are assigned a score. An alleged strength of MSS is that scoring would not be obvious to the rater (Katcher & Bartlett, 1979).

Examples of Mixed Standard Scales

An example of a Guttman scale development applied to an MSS application for performance appraisal is provided below for anchors with consistent combinations of high, medium, and low (Katcher & Bartlett, 1979).

MSS Error Counting System Anchors

Consistent Combinations:

High	Medium	Low
+	+	+
0	+	+
-	+	+
-	0	+
-	-	+
-	-	0
-	-	-

Before triads of anchor items for general performance areas (dimensions) are randomized, they are arranged in order from excellent performance to poor performance. The items are criterion-referenced to tasks instead of norm-referenced. Rosinger, Myers, Levy, Loar, Mohrman, and Stock (1982) present an example of a triad of anchor items. These items were identified for highway patrol troopers in Ohio for the general performance area of "stopping vehicles for a variety of violations."

- o "Stops vehicles for a variety of traffic and other violations.
- o Concentrates on speed violations, but stops vehicles for other violations also.
- o Concentrates on one or two kinds of violations and spends too little time on others."

MSS were established to reduce halo and leniency errors by mixing the statements. There is always the possibility that respondents will have difficulty identifying relevant behaviors, and matching the behavioral observation to the mixed item anchors (Katcher & Bartlett, 1979). This presents an ironic situation since MSS use could reduce two minor sources of error while introducing a source of error that had previously been controlled.

Comparisons of Mixed Standard Scales

MSS have been used to develop performance appraisal scales for police officers and highway patrol troopers (Rosinger, Myers, Levy, Loar, Mohrman, & Stock, 1982; Saal, 1979; Katcher & Barlett, 1979). The number of dimensions measured by MSS have ranged between 6 (Dickinson & Zellinger, 1980), and 10 (Katcher & Barlett, 1979). MSS scales are always anchored by triads (three items to describe proficiency levels). Then the items are randomized for rating. Items are either rated with a plus (+) minus (-), equal (=) or zero (0) for equal.

MSS are a variant of BARS and require only three anchor items for excellent, average, and poor performance. These anchor items tend to be shorter and more concise than those used with BARS. There is some evidence that MSS are similar in reliability and validity to BARS, and to graphic rating scales (Rosinger, Myers, Levy, Loar, Mohrman, & Stock, 1982; Saal, 1979). More research is required with MSS since Finley, Osborn, Dubin, and Jeanneret (1977) found the BARS format to be superior in convergent and discriminant validities to the MSS format.

In research performed by Rossinger, Myers, Levy, Loar, Mohrman, and Stock (1982), the majority of triads exceeded the .80 reproducibility level. Interrater reliability for the instrument as a whole was at the .90 level, and concurrent validity was .69 for the appraisal form as a whole. Dickinson and Zellinger (1980) obtained convergent and discriminant validity. The MSS format had as much discriminant validity as BARS and Likert formats. In research performed by Saal (1979), graphic rating scales were found to have greater interrater reliability than MSS. MSS had less halo error than graphic rating scales. The MSS investigated was a revised system for translating responses into numerical ratings. It was recommended that the revised MSS system would enhance rater acceptance and increase face validity. The revised system did not alter previous results obtained in the comparison of graphic rating scales and MSS.

The MSS format appears to perform psychometrically as well as other formats, e.g., graphic rating scale, Likert Scale, and BARS. As with the other formats, what seems to be important is the actual scale development. The MSS format appears to be as desirable as other formats in psychometric properties when developmental procedures are rigorous such as in the retranslation of expectations. Raters have not been as receptive to the MSS format, and have identified more preferred formats (BARS). Face validity and unidimensionality have also been issues with this scale. Most of the problems identified with the MSS appear to result from sophisticated attempts by researchers to remove minor sources of error and by concealing the scoring system from the rater.

Conclusions Regarding Mixed Standard Scales

MSS are structured using a Guttman rating method (Saal, 1979). Guttman scaling was developed as a response to deficiencies in the scaling techniques established by Thurstone and Likert (McIver and Carmines, 1981). Guttman scaling is designed to order subjects, as well as items, on an underlying cumulative dimension. The assumption is that a series of items in a Guttman scale belongs on a unidimensional continuum.

A high index of unidimensionality indicates that there are fewer inconsistencies in the rating of each item. It is imperative that unidimensionality be verified when the MSS are applied in field studies since this factor cannot be assumed. For example, in the evaluation of police sergeants and lieutenants, it was found that 95% of the ratings were inconsistent (Katcher & Bartlett, 1979). An inconsistent combination of ratings would result in a rating of "equal" or "the same" for the high anchor, and a rating of "not likely as good as" for the medium and low anchors.

There has been some concern regarding the MSS coding system since there are three possible responses to each behavioral statement. With three item anchors per dimension, there are 27 possible response combinations for any one dimension. MSS preclude the direct assignment of numerical ratings so that a coding system is required. The coding system generally used for MSS ratings does not appear to be internally consistent with face validity, although it is psychometrically consistent (Saal, 1979).

There is always the possibility that anchor items may be multidimensional instead of unidimensional, and this would yield inconsistent ratings. There is also the possibility that raters will inconsistently rate various behaviors while comparing the ratee to each anchor since the separate anchor may appear to represent different dimensions even if they are unidimensional. MSS may be more appropriate for use with cognitively complex raters (Schneier, 1977a) because of the potential problems with the item anchors. In a comparison of MSS to Likert scales and BARS, Dickinson and Zellinger (1980) found that raters preferred a BARS format. Rater acceptance can be an issue with MSS.

From a psychometric standpoint, MSS seem to be as sound as other scales when developmental procedures are thorough. MSS are not consistently superior when compared to other formats. Rater acceptance of MSS holds the potential for concern because of the inconsistent ratings obtained from multidimensional scales applied to field environments. Other areas of deficit have been where the anchor items were thought to be unidimensional, but did not prove to be. There is also the problem of the apparent lack of face validity for coding scores even though they are internally consistent. Last, but not least, there is the frustration of some raters not being able to identify anchor items since the anchor items are disguised by randomly mixing them.

CHAPTER IV

DESIGN OF QUESTIONNAIRE ITEMS

Questionnaire construction methods reviewed in this chapter focus on how to write questionnaire items and how to order the items for inclusion in the questionnaire. The importance of open and closed items, and when to use each type is examined. Guidelines are presented for how to word items, how many words to use in each item, and the influence of positive and negative wording. Research on the sequencing of items in a survey, and various approaches to balancing items are presented.

In the area of when to use an open item, and when to use a closed item, there have been many recommendations. However, much of what was written appears to be based on folk wisdom more than on empirical research. The literature does indicate that open-ended items are helpful in developing closed items and response alternatives, prior to the construction of a pretest.

It is known that the selection of wording in an item can change the response patterns to a significant degree. Even so, the state-of-the-art has not progressed to a point where researchers are able to consistently predict the effect of wording on item responses. It has been proposed that researchers may never be able to solve this dilemma because each time a word is changed in an item, there is the possibility that it will change the meaning of the entire item. One attempt to address this issue has been the creation of a system to guide the researcher in selecting words which go into items. The rationale is that it is possible to follow a set procedure to assist in identifying what words to select for items.

There have been some questions about not only what wording to include in an item, but also how many words to include in an item. The number of words to include in an item appears to be contingent on the content of an item. For most items (except threatening items), the number of words does not seem to influence results.

Once the actual items have been written, the researcher must decide how to order the items within the survey itself. This is another situation where researchers are cognizant of the fact that the order of items can influence the results, yet there is no known way to predict when item order effects will exist. Some researchers have suggested that randomly mixing the items will eliminate order effects. This does not appear to be a viable solution since some items won't make any sense to the respondent unless they follow a content-related sequence. The common advice for such a sequence has been to develop general items which are followed by content-specific items.

Balancing items so that they have positive or negative wording, or positive or negative response alternatives, was investigated for its influence on response effects. It appears as though items which measure personality traits are more influenced by balancing than items constructed for other applications.

Overall, the design of questionnaire items is tenuous since it is not possible to predict in advance the proper wording or ordering of items. Even so, this chapter provides some tentative recommendations to follow under the constraints of minimal empirical data.

4.1 OPEN-ENDED ITEMS AND CLOSED-END ITEMS

Description of Open-Ended Items and Closed-End Items

In addition to asking a question, the questionnaire designer also determines the amount of freedom the respondent will be given in expressing an answer to the question. A purely "open" item tells the respondents what topic to write about and provides blank space in which to write an answer. A purely "closed" item provides a set (closed, of course) of response alternatives and directs the respondent to select one of the response alternatives.

The terminology applied to these types of items may vary with the preference, research emphasis, or whim of the investigator. For instance, closed items have also been termed structured, fixed-choice, closed-fixed response, precoded questions, multiple-choice, forced-choice, rating scale, and this is by no means a complete list. Open items have been referred to as unstructured, free response, open-ended, essay, and even short answer.

The most popular questions with researchers have been the closed questions (see Section 2.1, Multiple-Choice Scales). Little research has been performed to substantiate the use of closed questions versus open questions, although the closed question is much easier to administer, score, and interpret.

Examples of Open-Ended Items and Closed-End Items

Cicchinelli, Harmon, and Keller (1982) conducted a cost effectiveness evaluation of three training devices for a portion of an avionics course at Lowry AFB. In addition to a troubleshooting test, they measured student and instructor attitudes toward the use of simulators and actual equipment in training. They also developed follow-up measures of training and job proficiency. In the assessment of field performance for avionics technicians, open and closed questions were both combined. Following is an example of how both types of questions can be combined, and the instructions accompanying the scale.

> "On the following pages, we would appreciate your help on this evaluation project. Please answer the questions to the best of your ability, using the graduated scale. The questions relate to your current working situation and your ATC training at Lowry AFB. Circle the point on the scale which most accurately reflects your situation or opinion."
>
> "Did your ATC training give you adequate training on the use of the patch panel as a troubleshooting instrument?"
>
> not at all — somewhat — very much
>
> |_____|_____|_____|_____|

"What aspects of your ATC training do you specifically use in your current field assignment?" _____

_____

_____

"What aspects of your ATC training do you use very little in your current field assignment?" _____

_____

_____

"What would you add to the overall ATC training program at Lowry AFB to better prepare avionics technicians for their field assignments?" _____

_____

_____

In a study comparing open versus closed questions, Bradburn and Sudman (1979) used a national sample from the National Opinion Research Center (NORC). Their questions started out with content focused on leisure and sport activities, and then transitioned to what would be considered threatening questions. They developed eight different questionnaire forms including open and closed questions. The open and closed questions were identical except that the closed questions incorporated response alternatives. An illustration of a question developed by Bradburn and Sudman (1979) includes an open and closed question in juxtaposition:

> "How would you describe your marriage, taking all things together? Would you say your marriage is completely happy, very happy, moderately happy, slightly happy, or not at all happy?"

Comparisons of Open-Ended Items and Closed-End Items

The Bradburn and Sudman (1979) research on open versus closed questions measured the following hypotheses: "H1 Open-ended questions elicit higher levels of reporting for threatening behavioral topics than closed-ended questions. H2 Long questions elicit higher reporting levels for threatening behavioral topics than short questions. H3 Familiar questions elicit higher reporting levels than questions employing standard research-chosen wording."

Bradburn and Sudman (1979) found that questions that have a "yes" or "no" response for behavior performed at least once do not support the three hypotheses listed above. For ("yes/no" response) questions about threatening behavioral topics, open-ended questions did not elicit higher levels of reporting than closed-end questions. Questions that ask the respondent to quantify the frequency or intensity of "sensitive" behavior produced different results. Hypothesis 3 continued to be rejected. Hypotheses 1 and 2 were supported for questions with threatening content. Open-ended questions thus are the preferred format for addressing threatening behavioral topics.

Schuman and Presser (1981) experimented with open versus closed questions, but did not focus on the area of threatening questions as Bradburn and Sudman had (1979). A work values experiment was conducted by Schuman

and Presser using an open and closed question format asking respondents what they most prefer on a job. They were not able to determine which type of question provided the most accurate view of respondent values. Almost 60% of the responses to the open question were not included in the fixed response alternatives in the closed question. These discrepancies may have been due to the fact that the fixed alternatives in the closed question may not have been pretested or that the response may no longer reflect current opinion. (Their fixed alternative had been previously generated by NORC Social Sciences Survey.)

Schuman and Presser (1981) hypothesized that the open question underestimated the respondents' perceptions of their concern regarding crime. The response category for crime and violence on the closed question had a percentage rate of more than twice that achieved for open responses identifying crime. An alternative hypothesis could also have been developed suggesting that the closed-end format may have induced overestimates by virtue of having presented fewer topics over which to distribute the responses. The open-ended interview is recommended as a way to discover response alternatives that the researcher did not think of.

Modification of the fixed alternatives in two following experiments by Schuman and Presser (1981) resulted in a shift in responses so that 58% of all the responses on the open form were included in the fixed alternative responses too. Previously it has been only 42%. Schuman and Presser considered the closed question form better to use than the open form since the responses are easier to code. Open question responses are not always that articulate, and responses can be vague.

Open questions are useful in pretesting questions to search out and select adequate response alternatives for closed questions. After question refinement is completed, closed questions appear to be superior for administration of the questionaire (Schuman & Presser, 1981; Orlich, 1978).

Conclusions Regarding Open-Ended Items and Closed-End Items

Because of the constraints involved in using open questions, most researchers have turned to closed questions for their surveys. Reservations about the use of open questions have been many. Some of the resistance to their use involves coding, tabulating, and quantifying the subjective responses -- this analysis can be extremely time consuming (Orlich, 1978). Open questions are also time consuming for the respondent (as well as the interviewer when interviews are conducted). For example, open questions have answers that are much more difficult to record than the closed questions. They either require more writing by a respondent or an interviewer depending on the type of questionnaire administration. Since open questions are more time consuming, this places a limitation on the number of questions that can be asked. It places an additional physical limitation on the questionnaire as to the number of pages and amount of space alloted for recording responses to each open question (Backstrom & Hurchur-Cesar, 1981).

There is a special role for the use of open questions. In the construction of a technically sound instrument, Schuman and Presser (1981) recommend conducting research with a large sample of the target population

by initially surveying the sample with open questions. The responses obtained are then transformed into response alternatives for closed questions. Backstrom and Hurchur-Cesar (1981) offer additional suggestions for the use of open questions. Open questions are able to elicit responses that can be used later in conjunction with quantified responses to add color to survey results. A qualitative analysis that includes anecdotal information can be included. Qualitative analyses compare the data collected from the open-ended questions with some predetermined standard of what it should be. Qualitative analyses are theoretical, and not quantified. Open questions are also a way to explore a respondent's attitudes and in-depth motivations.

Bradburn and Sudman (1979) found long open questions to be most useful in obtaining information from respondents under specific conditions. These questions were directed toward gaining information about sensitive behavior (gambling, alcohol, drugs, sexual activity, and income) using familiar wording. Differences between open questions and closed questions for threatening topics were significant at the .05 level. Threatening questions which request information about whether a behavior took place, and only require a "yes" or "no" answer, obtain the same response whether they are open or closed questions. Bradburn and Sudman indicated that it may be easier for respondents to acknowledge they were involved in a behavior than to indicate their degree of participation.

The research conducted by Schuman and Presser (1981) suggests that the differences in responses to open and closed questions may be differential across populations. Apparently more educated populations tend to have greater congruity between open-ended and closed-end forms, while less educated respondents have more divergence between these forms. This disparity may result from the lower motivation possessed by the less educated to write essay answers.

There is a need for both open-ended and closed-end questions. Open questions are most appropriately used for pilot testing prior to selecting a closed question response format. Open-ended questions may be useful when researching sensitive content areas that might be perceived as threatening by respondents. Research that compares open and closed questions has been sparse, although this topic has been under consideration for over 50 years. Therefore, the conclusions rendered here are somewhat tentative. It has been the standard operating procedure for most researchers to use closed questions as the primary type of question in their refinement of survey instruments.

4.2 WORDING OF ITEMS AND TONE OF WORDING

Description of Wording of Items and Tone of Wording

There have been a number of investigations regarding what is the best way to word items in questionnaires. Application of the wording of items has been diverse and includes: questionnaires used for surveys, questionnaires used for performance rating scale items, and questions for test items. Many of these investigations have followed the armchair philosophy approach to science by coming up with commonsense advice on how to word items. There have been some empirical investigations (experimental designs using quantitative methods) for the wording of items.

Some of the research has focused on wording items by developing a dichotomy of positive or negative statements (Ory, 1982). Positive or negative wording of items was explored to determine whether respondents would have a tendency to endorse positively worded items and reject negatively worded items (Deaton, Glasnapp, & Poggio, 1980). Other kinds of dichotomies have also been proposed for wording items. For example, Orlich (1978) suggested that questionnaire items could be worded so that they are either personal or impersonal. Supposedly, items worded personally will be more personal than they would be if worded impersonally. There is the potential that the personally worded item will be more specific to the experience of the respondent. This may provide the researcher with results that have greater accuracy for items that are non-threatening. For threatening items written in personal terms, there may be a tendency to underestimate a behavior which would result in less accuracy. Of course, it also is possible to include both personal and impersonal versions of items in the same questionnaire.

Researchers who are responsible for the wording of items face many problems since it is known that a slight change in wording could change the results of the survey (Orlich, 1978). A potential pitfall for wording items has been identified. The use of technical words and technical jargon would be understood by professionals, but may not be understood by respondents (Labaw, 1980; Strang, 1980; Backstrom & Hurchur-Cesar, 1981). Some words embedded in questionnaire items cause ambiguity for respondents. This ambiguity may be created for a number of reasons. An illustration of this would be words to which a respondent cannot relate. This could be caused by words which lie outside their experience (Backstrom & Hurchur-Cesar, 1981). Other reasons for ambiguity may be the use of words embodying complicated abstractions or words that have multiple meanings (Labaw, 1980).

Backstrom and Hurchur-Cesar (1981) indicated that each word needs to be viewed not only for its own meaning, but also by the context in which the word is found. Items may be distorted by emotionally charged words or by using terminology which indicates to the respondent that one alternative may be more desirable than another (loading the question). Each item needs to be worded so that the meaning will be clear and unequivocal to all respondents. Individuals who write questionnaire items should screen for words which would cause a biasing of results. Most blatantly biasing words

probably are identified and removed from survey items. It is difficult to predict in advance which words will bias an item. Schuman and Presser (1981) found that it was not the blatantly biasing words that cause the most distortion, but the much more subtle words. They felt that the blatant words were so outstanding compared to the other words that it was impossible not to notice these biasing words in an item.

Examples of Wording of Items and Tone of Wording

In some sample questions developed by Orlich (1978), the differences in item content for personally written items versus impersonally written items were illustrated. Orlich developed items regarding interpersonal relationships with managers. Personally oriented items requested respondents to rate their relationships with management. Items which were impersonal requested respondents to rate how other employees get along with individuals on the job, and how work is rated by managers.

Smith (1981) presented examples of questions that were highly ambiguous. Apparently, many of the respondents did not consider the first item in a sequence of questions in a literal sense. Instead, the respondents did not seem to be able to imagine how the consequences of their first answer would impinge on their responses to the following items. Below is one set of items that obtained many illogical response patterns due to the ambiguity of the wording (Smith, 1981).

"Are there any situations you can imagine in which you would approve of a policeman striking an adult male citizen?" YES, NO, NOT SURE

"Would you approve if the citizen . . ."

A. "had said vulgar and obscene things to a policeman?" YES, NO, NOT SURE

B. "was being questioned as a suspect in a murder case?" YES, NO, NOT SURE

C. "was attempting to escape from custody?" YES, NO, NOT SURE

D. "was attacking the policeman with his fists?" YES, NO, NOT SURE

For all the respondents who said "no" to the first question, 86% selected "yes" to one or more of the latter items (A, B, C, or D). Additional structuring of these questions could have been provided to alleviate the ambiguity which resulted.

Schuman and Presser (1981) reported on the work of Mueller (1973) where Mueller researched the Korean and Vietnam wars regarding public opinion data. A trend was identified by Mueller in an experiment using a Gallup question. When questionnaire items mentioned the threat of Communism, support for U.S. intervention was increased. The original item used

in a Gallup Opinion Index in 1967 was later used in an experiment by Schuman and Presser (1981) along with a modified item that incorporated the threat of Communism. They found that support for U.S. military intervention can be increased by as much as 15% if an item incorporates the possibility of a Communist threat.

Usually, the blatant attempts to bias an item by tone of wording are not so likely to succeed. In addition, not every change in wording will create a significant difference among marginals (Schuman & Presser, 1981). Marginals are percentages of responses to each response alternative for each item in a questionnaire. Schuman and Presser reported the work of Stouffer (1955) where an item identified individuals who were against churches and religion as being bad and dangerous. This blatant language appeared to have no effect on the responses.

Comparisons of Wording of Items and Tone of Wording

A great deal of the literature on item wording and tone of wording does not fit into the framework of an experimental design. Many recommendations for the way in which items are worded are based on the actual field experience (folk wisdom) of individuals who design questionnaires. These researchers' recommendations are more or less consistent across the literature. For example, individuals who design questionnaires would agree that the use of ambiguous words in an item would distort the intent of the item. The meaning of the item would then be ambiguous to the respondent(s) (Backstrom & Hurchur-Cesar, 1981; Smith, 1981; Labaw, 1980; Orlich, 1978).

Ory (1982) investigated the positive and negative wording of questionnaire items. These items were embedded in a performance evaluation scale. Ory hypothesized that respondents would be influenced by positively worded items and by negatively worded items. The results of two studies conducted by Ory indicated that the positive and negative wording of the items did not affect the respondents. There were no significant differences found for rating items with positive or negative wording. Research performed by Deaton, Glasnapp, and Poggio (1980) indicated that positively worded items received higher mean responses than negatively worded items. This trend in rating positive items higher, and negative items lower, did not reach statistical significance. Respondents appeared to express a preference for or agreement with positively worded items by rating them higher than negatively worded items. Deaton et al. provide limited evidence that the tone of wording (positive or negative) can influence response patterns.

Schuman and Presser (1981) hypothesized that respondents with strong attitudes toward a topic would be less influenced by the tone of wording in a survey item, and that respondents who did not have a strong attitude toward the content area would be more easily influenced by the tone of wording in an item. They were not able to establish convincing evidence to support their hypothesis.

Items where respondents frequently ignored the absolute phrasing were the focus of research conducted by Smith (1981) (see Section 4.2, Examples of Wording of Items and Tone of Wording -- policeman striking citizen). The wording on the survey items Smith used did not prevent respondents from answering the questions with contradicting response patterns. Respondents

who answered the first item in a series of 5 items with a "no" would also have to answer "no" to the other 4 items in order to maintain a logical sequence. However, 86% of the respondents who answered "no" to the first item of the series provided a contradictory response to the rest of the series. Smith's investigation of incongruity for ambiguous item response concluded in a profile for those particular respondents. The respondents who answered "no" to the first question regarding the approval for a policeman to strike an adult male and then answered "yes" to one or more items approving such striking were investigated further. Additional data obtained from these respondents was: (1) Interviewer's assessment of the respondent's comprehension, and a 10-item word identification test measuring verbal ability and years of schooling; (2) Respondents were asked about their attitudes toward the judicial system and questions about first-hand experience with varying degrees of violence; and (3) The respondent's propensity to check "don't know" response alternatives was examined. Respondents with contradictory response patterns were non-white, had less education, less verbal achievement, and lower comprehension than other respondents. These respondents were more likely to be female, and their attitude was in favor of the initial statement in each series of five items.

Schaefer, Bavelas, and Bavelas (1980) developed a method to ensure that respondents would only be subjected to items that they could understand. The technique that they used is called "Echo." They developed items that were used in a performance rating scale. It would be possible to use the "Echo" technique in the development of survey items too. Essentially, the "Echo" technique is a method for wording questionnaire items in the language of the respondents. A detailed procedure for using the "Echo" technique is available from J. B. Bavelas (1980).

The "Echo" technique assumes that there are two separate populations in the development of questionnaire items. One population is the researchers, and the other population is the respondents. Phrasing of items needs to be in the language of the respondents, and it requires content validation. A summary of the "Echo" technique includes the development of a pool of items generated by a survey directed to the target population. The sample of potential respondents from the target population follows printed guidelines to write the items. Another sample from the target population is selected to sort items into categories. Part of this process includes concurrence by the members of the sample that the categories are mutually exclusive.

Schaefer, Bavelas, and Bavelas (1980) determined that a questionnaire constructed by the "Echo" method was rated by respondents as superior to four other questionnaires at the .001 level of significance. The results they obtained support a suggestion made by Labaw (1980). Respondents can explain what they mean to assist researchers in clarifying item wording. This is a way to assure that questionnaire items do not become instruments to force researchers' language, jargon, and values upon the respondents.

Conclusions Regarding Wording of Items and Tone of Wording

Researchers are cognizant of the fact that the wording of an item and/or the tone of wording has the potential to change the marginal responses to a significant degree. Yet, being able to predict when this effect will take place, and by what kind of words, seems to be beyond the capability of research at this time. This is not to say that in some instances researchers have not been able to predict the effect of wording in items (Mueller, 1973; Deaton, Glasnapp, & Poggio, 1980). The results aren't consistently replicable.

Schaefer, Bavelas and Bavelas (1980) pointed out one of the primary factors inhibiting research for the identification of words for inclusion in items. A standardized set of acceptable words or standardized questionnaires may not be what is needed for writing reliable and valid items. There are too many contexts for word inclusion in items, too many different populations to address, etc. What may be needed is a procedure or method to identify specific words to be used in items, and the structure of the item itself. Obviously, such an approach calls for greater rigor, time, and work by the research community. The selection of words for inclusion in items must be contingent on respondent experience with the content. The only way to ensure that respondents will understand the wording is to use the language of the respondents. Currently, there are no clear-cut ways to control for word bias with the exception that questionnaire item designers be sensitive to the issues of bias. If a word is so outstanding that there is no doubt that it would bias an item, then there is a good chance that the reverse will take place (Schuman & Presser, 1981). If it is that noticeable, then respondents would probably not be influenced by the biasing words either.

4.3 LENGTH OF ITEMS AND NUMBER OF ITEMS

Description of Length of Items and Number of Items

In the construction of a questionnaire, the issue of length may be addressed from many perspectives. For example, length could mean: the number of pages included in a survey, the number of items used in a questionnaire, or the number of words employed in each item. In an educational survey conducted by Layne and Thompson (1981), they focused on the number of pages in a survey. The number of items was held constant regardless of the number of pages. Bradburn and Sudman (1979) compared long and short items. They defined longer questions as exceeding 30 words. Their research was applied to a national survey sample. Mullins, Earles, and Wilbourn (1979) compared performance appraisal items for optimum number. These items were incorporated into a rating form for non-commissioned officers (NCOs) participating in Air Force seminar groups. Across all instructional applications, the issue of length of items, number of items, etc. must be addressed each time an instrument is devised. However, research in this content area has been diverse and limited.

Examples of Length of Items and Number of Items

Mullins, Earles, and Wilbourn (1979) hypothesized that when raters are not trained, they will rate performance only on a general concept of excellence. They felt that requiring the raters to assess individuals on many separate characteristics would not improve the accuracy in their ratings. They designed instruments with varying numbers of items to investigate this hypothesis (5, 10, and 20 items). An illustration is provided below of their 20-item rating scale.

	Below Average	Average	Above Average	Well Above Average	Outstanding
1. "Learning Ability - acquires knowledge accurately and quickly"	(A)	(B)	(C)	(D)	(E)
2. "Leadership - effectiveness in getting ideas accepted and in guiding others to accomplish a task"	(A)	(B)	(C)	(D)	(E)
3. "Quality of Work - produces work of high quality"	(A)	(B)	(C)	(D)	(E)
4. "Motivation - strong desires to accomplish goals and objectives"	(A)	(B)	(C)	(D)	(E)

		Below Average	Average	Above Average	Well Above Average	Out-standing
5.	"Follows Instructions - follows directions as prescribed"	(A)	(B)	(C)	(D)	(E)
6.	"Bearing and Behavior - maintains professional conduct and appearance"	(A)	(B)	(C)	(D)	(E)
7.	"Accuracy - precision and carefulness in work performance"	(A)	(B)	(C)	(D)	(E)
8.	"Oral Communication - expresses ideas clearly, logically, and grammatically in conversation"	(A)	(B)	(C)	(D)	(E)
9.	"Problem Analysis - identifies and analyzes problems which require action"	(A)	(B)	(C)	(D)	(E)
10.	"Initiative - self-starting, rarely needs a push to get going"	(A)	(B)	(C)	(D)	(E)
11.	"Quantity of Work - accomplishes a large amount of work"	(A)	(B)	(C)	(D)	(E)
12.	"Written Communication - expresses ideas clearly in writing with good grammatical form"	(A)	(B)	(C)	(D)	(E)
13.	"Punctuality - prompt in keeping engagements"	(A)	(B)	(C)	(D)	(E)
14.	"Adaptability - changes attitude and behavior to meet the demands of the situation"	(A)	(B)	(C)	(D)	(E)
15.	"Dependability - does assigned tasks conscientiously without close supervision"	(A)	(B)	(C)	(D)	(E)

	Below Average	Average	Above Average	Well Above Average	Out-standing
16. "Emotional Stability - stability and calmness under pressure and opposition"	(A)	(B)	(C)	(D)	(E)
17. "Human Relations - gets along well with fellow workers and works effectively with them"	(A)	(B)	(C)	(D)	(E)
18. "Judgment - makes good decisions among competing alternatives"	(A)	(B)	(C)	(D)	(E)
19. "Knowledge of Duties - understands the requirements for effective work performance"	(A)	(B)	(C)	(D)	(E)
20. "Honesty - straightforward and truthful in dealing with others"	(A)	(B)	(C)	(D)	(E)

To obtain higher reporting levels by respondents when threatening questions are asked about their behavior, Bradburn and Sudman (1979) found that longer items were best. Items with 30 or more words achieved best results while items with fewer words (less than 30) did not elicit reporting levels which were as high. One of the longer items developed by Bradburn and Sudman had 49 words, and the content was about the use of drugs. A threatening question developed by Sheehy (1981) is illustrated below. This question had over 100 words. It was included in a Life History Questionnaire that was completed by 60,000 respondents.

> "Below is another chart, similar to the one you have just completed. Complete this one in the same manner. For each age period you have lived through, place the number(s) of the one or two most important feelings, changes, or experiences in the appropriate boxes. (This time the list includes 15 items.) You may use each number as many times as you like. Then consider each of the periods you have yet to live through. For each future period, place the number(s) of the one to two most important feelings, changes, or experiences that you think are likely to occur during each of those periods."
>
> 1. "Felt that time was running out"
> 2. "Felt this was my last chance to 'pull away from the pack'"
> 3. "Felt confused or conflicted about choice of career or career direction"
> 4. "Seriously questioned my parents' beliefs and values"
> 5. "Felt stagnant in my work"

6. "Felt stagnant in my home life"
7. "Felt truly middle-aged"
8. "Felt I had probably reached my peak earning years"
9. "Asked myself, 'Is anything worthwhile? Does anything matter?'"
10. "Felt 'no longer young'"
11. "Suddenly noticed my friends were looking old"
12. "Had serious marital difficulties"
13. "Felt confused or conflicted about proper sexual standards for myself"
14. "Began to think seriously about my own mortality"
15. "Became seriously depressed or discontent"

	Age Periods					
Experience	18-28	29-35	36-45	46-55	56-65	66 +
Numbers	☐☐	☐☐	☐☐	☐☐	☐☐	☐☐

In research performed by Layne and Thompson (1981), they held the number of items constant, but expanded the number of pages from one to three on their survey instruments. (Their short form consisted of 30 items on one page, and the long form consisted of 30 items spread out over three pages with 10 items to a page.)

The perception by investigators as to what constitutes length when designing items, as well as designing entire questionnaires, is quite diverse. For example, how long is a long item? Is it more than 17 words or is it more than 30 words, etc.? How long is a long questionnaire? Does a long questionnaire have 20 items or does it have 80 items? Does a long questionnaire mean number of pages in length instead of number of items in length? Of course, there are no definitive answers to these questions since each researcher defines what they believe is short or long for number of words in an item, number of items in a questionnaire, and number of pages used for the questionnaire.

Comparisons of Length of Items and Numbers of Items

Research in this area is diverse, but limited, so that actually generalizing from one study to another has not been possible. The subjects used in the reported studies encompass NCOs from Air Force seminar groups, a national sample of adults, and Master of Education graduates.

In 1981, Layne and Thompson reported on their research survey on 400 Masters in Education graduates to investigate the influence of follow-up letters. They analyzed the return rate for short and long forms (1-page versus a 3-page format) when the number of items is held constant. They determined that questionnaire length (number of pages) and use of a follow-up letter were not related to response rate. Increasing response rates through the use of an abbreviated survey (fewer pages) could not be supported based on the results of this study.

Bradburn and Sudman (1979) compared: open-ended and closed-end questions, long and short questions, and familiar-worded and standard-worded

questions for a national sample of adults. They defined long questions as those using more than 30 words. The hypothesis was that more information would be obtained for responses to threatening questions with more words. They found that there was no format difference for the responses to threatening questions. This finding was for questions that requested information on whether a behavior was performed only once. These are questions which required only a "yes" or "no" response. When questions are structured to obtain the frequency about a "sensitive" behavior (in this study a sensitive behavior had to do with drug use, sexuality, alcohol consumption, etc.), items of greater length (more than 30 words) tend to increase the reporting. They suggested that non-threatening types of questions are not affected by the number of words in the item.

The most efficient number of items to include in performance appraisal rating instruments was investigated by Mullins, Earles, and Wilbourn (1979). Their subjects were 132 Air Force NCOs assigned to the Air Training Command. Subjects rated peers on 5, 10, and 20 item instruments. They hypothesized that they would rate performance on a general concept of excellence. They felt that adding additional items to the rating form would not influence the raters ability to discriminate. They concluded that more than five items on an instrument designed to measure performance was not advantageous. In this particular study, subjects were later asked to identify peer profiles based on peer ratings. More than five items did not add to the accuracy of the ratings when peer profiles were used as a criterion.

Conclusions Regarding Length of Items and Number of Items

Research in this area is diverse and limited. It is not possible to generalize any specific theories or models about how many items to include in a questionnaire or about how many words to include when writing an item. From the limited data presented, there was an indication that the number of pages used in a questionnaire did not influence response rate when the number of items was held constant (Layne & Thompson, 1981). When Mullins, Earles, and Wilbourn (1979) compared number of items (5, 10, and 20) to use in rating performance, they found that five items were adequate in their scale construction. Questionnaires constructed with a large number of items may not provide any more valid measurement than questionnaires constructed with a smaller number of items. Their study employed a single external criterion of class standing to compare the ratings against. Perhaps the raters were unable to differentiate between items (traits), and were reflecting their general perception of the ratee's performance. Even if the respondents were better able to differentiate, item reduction techniques are recommended to reduce the number of items used in a questionnaire. Item reduction is a common technique used in the development of questionnaires. It has been used extensively in the development of behaviorally anchored scales and in marketing surveys.

Bradburn and Sudman (1979) researched threatening questions. Format differences did not influence respondents' willingness to report the occurrence of the behavior. Measurement of the frequency of a behavior was best achieved through the use of open-ended questions which had 30 or more words. Apparently, responses to non-threatening questions are not influenced by number of words in a question. This finding for non-threatening questions is consistent with research findings in Section 7.1, Questionnaire Layout.

4.4 ORDER OF ITEMS

Description of Order of Items

The order of items may be configured in many ways, dependent on how the items will be used. For instance, when items are used in opinionnaires, investigators sometimes ask multiple questions on a topic. This may reveal a greater depth of information as the questions become more specific and continue in a sequence. Respondents try to be consistent in this type of situation. However, it is possible that the respondents' answers are based on information they are obtaining by reading the previous questions. The responses may not be well thought-out attitudes on the topic (Labaw, 1980).

Schuman and Presser (1981) found that initial items may influence later items. Items which are replicated in different contexts may not control for order effects, but may be confounding order effects with true change. It was determined that general items are more prone to order effects than more specific items.

Item ordering for test construction has been investigated for writing items in an easy to hard sequence. Items which are found in tests are sometimes ordered by the degree of difficulty. Easier items are presented first followed by succeedingly harder items. The easy to hard sequence found in constructing items uses the rationale that if individuals do not answer an item correctly then they will probably get the next item incorrect too. If they get an item correct, there is a better chance of getting the following item correct.

Examples of Order of Items

Labaw (1980) concluded accepting responses at face value for initial items may not provide the researcher with valid results. For example, what party a person voted for in a previous election (Democrat or Republican) is a better indicator of future voting behavior than responses that indicate the respondent prefers to vote for "the best candidate." Labaw provides an example of item ordering which sorts out this type of inconsistency:

1) "I vote for the man, not the party."
2) "What are the characteristics of the man you vote for?"
   Answer: "Honesty."
3) "How do you define honesty?"
   Answer: "An honest man is one who votes on my side of the issues."
4) "How do you know he votes on your side of the issues?"
   Answer: "Because he is a Democrat."

In a study on the effects of item order, McFarland (1981) investigated whether general items on a survey should be followed by items which are more specific. One of the general items pertaining to energy requests the respondent to describe the current energy problem in the U.S. The respondent is to rate it in a range between "extremely serious" and "not serious

at all." Specific items focused on specific attitudes toward energy related content areas, such as: causes of the gasoline shortage, windfall profits tax on oil companies, nuclear energy, and strip mining regulations that had the potential to increase coal costs.

Comparisons of Order of Items

Surveys usually consist of consecutive items which are related by topic. The ordering of items for context effects occurs when two or more items are presented together on the same topic or with closely related topics. Items which are general and not specific may be prone to context effects. Yet, the meaning of the items would be changed if they were separated from their topic areas (Schuman & Presser, 1981). The current state-of-the-art for context effects suggests that all items which are interrelated by content area may be affected by context effects. There is currently no way to predict which items will have context effects.

The ordering of items has not usually been subjected to experimental research. Some investigators tend to give prescriptive advice on the way to sequence items in a survey (from general to specific in topic areas). McFarland (1981) examined general and specific survey items for order effect. No significant relationship was found between order effects, sex, and education. Order of the items did not appear to effect the relationships between the general and specific items. However, 2 out of 17 relationships did reach significance at the .02 level. It is proposed that a stronger survey instrument may be provided by designing general items first, followed by specific items on related topic areas.

Another approach to dealing with content related item ordering was proposed by Labaw (1980). Using this approach, each item is formulated to follow a logical progression. This may provide a better opportunity to have the responses screened. The respondents can be assessed for their knowledge and understanding of the topic area to legitimately answer the item. There is certainly no guarantee even then as to the respondent knowledge base.

The issues related to order of items have been investigated by a number of researchers (Spies-Wood, 1980; Dambrot, 1980; Gerow, 1980; Schmitt & Scheirer, 1977; Spiers & Pihl, 1976) for multiple-choice questions. The question of what is the right order of items has focused on including items in a sequence where the items start out easy and then become hard.

Overall, respondent attitude toward success in responding to an item seems to have an effect on the test performance. Sequencing easy to hard items assists respondents in building up a feeling of success according to Spies-Wood (1980). Dambrot (1980) found that sequencing items from easy to hard had little effect on respondent performance. Dambrot also reported the work of Schmitt and Scheirer (1977) and Spiers and Pihl (1976), where the item order did not have a demonstrable effect on respondent performance. Gerow (1980) found no significant difference for the ordering of easy to hard items versus random ordering on test construction and administration. The weight of the evidence does not appear to support the proposition that ordering items from easy to hard facilitates questionnaire-answering performance.

Conclusions Regarding Order of Items

Questionnaires are plagued by contextual effects attributed to item ordering. This occurs when a number of items are developed on the same topic and then grouped together by content. The result of this type of item composition differs by questionnaire. Consistency of responses across items may emphasize a perceived similarity or it may have the opposite effect where differences are emphasized. Apparently, contextual effects can be minimized by generating items which are more specific in content (Schuman & Presser, 1981; McFarland, 1981).

The quality of responses to items on a questionnaire will be determined by the respondent's background and knowledge of the topic area. A series of specific items (versus general items) will provide information about whether the respondent understands the content of the items. It should expose any logical inconsistencies in response patterns. Respondents with limited or no experience regarding the content area may deviate from the original approach. Their answers to questions change as they become more familiar with the topic through order effects. Researchers may not want to accept early responses as having face validity. Order of items and consistency in logic can be reviewed in a pretest by questioning respondents on what they think each item means (Labaw, 1980). While additional research is needed on the effects of ordering multiple-choice items from easy to hard, the results of the research performed so far indicates that random ordering produces results no different from easy to hard ordering.

It is assumed that item order effects exist, yet it is not possible to predict when they will occur. Some research has indicated item order effects in marginals. Marginals are percentages of responses to each response alternative for each item in a questionnaire. This distribution is considered a function of the wording of the item or possibly the ordering of an item. The wording of items has been known to change the size and/or the direction of relationships for the distribution of responses to the response alternatives. The differences in percentages attributed to each response alternative is studied for items. Research designs have been developed to compare the ordering of items, and to compare the wording of items by assessing the differences among the marginals. Apparently, it is possible to have order effects without their being displayed in the marginals. Order effects also have been measured by finding correlations among items which have been affected.

4.5 BALANCED ITEMS

Description of Balanced Items

Some investigators strive to select scale items which consist of items that are positively and negatively worded. Their intention is to create a "balanced" scale. When the items have been exposed to normal scale reduction procedures, they need to retain their construct validity (Ray, 1982). The decision to balance scale items has usually followed from the research findings where acquiescence response tendencies were identified. Investigators wishing to avoid a response set have used balanced items for this purpose. Ory (1982) provided an illustration of this effect from the research of Cronbach (1946, 1950) and Couch and Keniston (1960). Cronbach (1946, 1950) obtained results where respondents used positive response sets, and Couch and Keniston (1960) determined that respondents had a tendency to use a positive or a negative response set.

It has been suggested that some respondents mark scales according to their propensity to select alternatives along either the positive or negative continuum. Ory (1982) hypothesized that to avoid this effect, two types of questionnaire items should be used. Some should be positive and others negative in orientation. It was hoped that this would balance out the scale. Ory used negative or positive wording of items to measure this phenomenon in two experiments. The placement and wording of items was investigated. Different forms were developed that included varying numbers of negatively worded questions. One form had only positively worded items and no negatively worded items. Other forms had 10, 20, or 30 scale items which were negatively worded. One of the forms had only negatively worded items. All questionnaires contained 30 items each. Both studies indicated that the positive or negative wording did not significantly influence the results.

Examples of Balanced Items

Individuals who mark bipolar scales (such as the Minnesota Multiphasic Personality Inventory) to measure personality traits may have been responding in ways to enhance their own social desirability. This rating tendency has produced the confounding of the socially desirable response with the trait being rated (Klockars, 1979). To avoid this type of response set, Klockars developed a modified approach to the construction of bipolar scales by anchoring with adjectives. Respondents were provided with scales that had only one endpoint. Klockars provided an example of the endpoints as follows:

Hot ____ Cool

Hot __X__ Cold

Klockars' subjects were to select between two adjectives which would be used as the other endpoint. The choice consisted of a positive endpoint that was considered socially desirable or a negative endpoint that was considered socially undesirable. Socially desirable endpoints that were in opposition to the question stem were thought to confound trait and desirability.

Johnson (1981) examined response styles for the order of presentation of positive and negative items at the first position/endpoint in semantic differential scales. The sample included male readers of Horizons USA in Great Britain, Italy, Phillipines, and Venezuela. The semantic differential scale consisted of 11 intervals identified as 0-10. Two versions of the survey questionnaire were developed. One scale had positive anchors first and the second scale had negative anchors first. An illustration is provided below listing the positive and negative anchors Johnson used for the item presentation in the 11 bipolar scales in four countries:

Item

Accurate-
Inaccurate
Authoritative-
Not authoritative
Impartial-
Prejudiced
Well intentioned-
Questionable intentions
Timely-
Old, dated
Important to me-
Unimportant to me
Thought provoking-
Bland
Relevant to my interests-
Irrelevant to my interests
Visually attractive-
Visually unattractive
Credible-
Not credible
Best magazine of its kind-
Worst magazine of its kind

There was no significant difference between the two formats for the presentation of positive or negative anchors placement on the scale. Order of presentation was not associated with response style across multinational settings (Johnson, 1981).

Ory (1982) used items from the Instructor and Course Evaluation System (ICES) to study the effect of negatively worded items on respondent ratings. Ory's research indicated that the positive or negative wording did not significantly influence the results. An example is presented here of the positive and negative items Ory included in his questionnaire. Students rated their course and instructor on a 5-point scale with anchor alternatives from "poor" (=1) through "excellent" (=5).

"Positive version = Exams covered a reasonable amount of material"

"Negative version = Exams covered an unreasonable amount of material."

Balancing questions in attitude surveys has also consisted of an approach termed "formal balance." Some researchers have tried to persuade

their respondents that it is perfectly acceptable to select both positive and negative response alternatives. One way of doing this is to use items that contain both positive and negative content. These types of survey items are considered to have "formal balance" (Schuman & Presser, 1981).

The researchers at the Army Research Institute, Fort Hood, recommend avoiding the use of unbalanced directionality or intensity of attitude in the stem of a question. They usually work with rating scales similar to the semantic differential, which simplifies the composition of the stem. These researchers do not request a rating for how effective a system is, but instead they ask for a rating of how "effective-ineffective" the system is. Alternatively, they delete the dimension from the stem altogether, and show the respondent the dimension only in the list of response alternatives. This approach is thought to create a formal balance in the response alternatives. Using these techniques, the stems either have a formal balance or avoid specifying the dimensionality of the rating.

Balance in questionnaires has been achieved in diverse ways for different applications. Professional survey organizations use internal methods to balance questionnaire items by balancing wording within each item to include positive and negative statements. Questionnaires used for student rating forms have contrasted positive items with negative items. Marketing surveyors have anchored endpoints in semantic differential scales with positive endpoints first or negative endpoints first. Balancing has been used to anchor endpoints for personality measurements as a way to control for socially desirable response sets.

Comparisons of Balanced Items

Positively and negatively worded items were developed to balance a Likert scale constructed to measure environmentalist attitudes (Ray, 1982). Four questionnaires were ultimately developed. Two questionnaires were balanced with 12 items and 20 items, and two questionnaires were not balanced. They contained 12 items and 20 items also. Ray was interested in determining whether the construct validity of the scale could be maintained during item reduction procedures commonly used in scale development. These four questionnaires were correlated with the initial 77-item scale and with each other. Correlation coefficients ranged between .78 and .87 for reliability. For validity, correlation coefficients ranged between .80 and .90. Normal scale reduction procedures did not jeopardize initial and final forms of a balanced scale or an unbalanced scale. Construct validity was maintained when forced balancing was used. This research was performed through New South Wales University in Australia. Seventy-five respondents were involved in this study.

Using a semantic differential scale (with 11 intervals), Johnson (1981) investigated the presentation of either positive or negative endpoints displayed first at the left-hand side of the scale. Johnson was concerned with the possibility of a response set where a respondent consistently marks a positive or negative stimulus anchored word depending on its placement on a bipolar scale. Primarily male subjects were selected from Great Britain, Italy, Phillipines, and Venezuela on the basis of their readership of Horizons USA magazine. The type of response style focused on

in this study is the tendency to consistently answer positively or negatively. This tendency depends on the placement of the positive or negative endpoint displayed at the left-hand side of the semantic differential scale. When the data was combined for all four countries, there was no significant difference between the two formats. The order of presentation for the placement of positive or negative endpoints was not associated with response style since there was no clear pattern across the individual dimensions. However, when the data was analyzed on a country-by-country basis (instead of combined for all four countries), there was some evidence that response styles differ nationally. For the Phillipines there was a bias toward positive stimulus words, and for Italy there was a bias toward negative stimulus words.

Klockars (1979) researched semantic differential scales for response sets (socially desirable responses) that were confounded with trait self-descriptions on clinical instruments. The results indicated at the .05 level of significance that subjects confound the desirability dimension with the trait dimension. Klockars found that the presentation of a negative adjective (one that was socially undesirable) would influence the selection of a positive adjective (opposite in meaning).

Schuman and Presser (1981) established balanced questions by balancing the pro and con response totally in one question. For example, on a question regarding unions, the balanced survey item was constructed as follows: "If there is a union at a particular company or business, did you think that all workers there should be required to be union members, or are you opposed to this?" They investigated whether "balancing" items this way would change survey results in comparison to items which were not balanced. They conducted four experiments with three of the experiments giving no indication of a difference. Only the fourth experiment showed significance, with a 9% increase for the balanced item in the negative direction. They were not able to obtain evidence to substantiate that balanced items affect response on attitude surveys (there appeared to be no difference in distribution). In other research performed by Schuman and Presser (1981), they found that adding a counter-argument into an item did not serve to balance the item. Instead, it established a new item which influenced the negative response.

Ory (1982) investigated whether positively or negatively worded diagnostic items would influence response sets for global items used in the evaluation of instruction. Diagnostic items were defined by Ory as items which "...measure student judgments and observations of specific behaviors of the instructor, instructional techniques, and detailed student outcomes." Global items were defined as items which "...measure student evaluations of general areas of instruction." Ory determined that the positive or negative wording of the diagnostic items did not influence the results. In another attempt to measure effects of positive and negative wording of items, Deaton, Glasnapp, and Poggio (1980) compared forced-choice scale items for positive or negative wording, item length, and effects of vague adverbs used to modify sentences, such as: "I 'sometimes' enjoy being outdoors." The main effects for item direction (positive or negative wording) and item length were significant at the .05 level, although none of the interactions were significant, nor was the main effect for modifier intensity. Apparently longer items produced responses that

were closer to the center of the response scale. Shorter items yielded more positive responses. Items that were positively worded received higher mean responses than negatively worded responses.

Conclusions Regarding Balanced Items

From the research presented regarding balanced questionnaires, it can be seen that the term "balancing" means different things to different researchers. In two studies (Johnson, 1981; Klockars, 1979), the balancing of anchors was investigated. The Johnson scale was used in a cross national survey, and the Klockars scale was used for clinical purposes to measure personality traits. The manipulation of anchors to achieve balance for these two semantic differential scales resulted in different conclusions. Balancing positive and negative anchors did not indicate a response set overall across four conditions. When anchors were balanced on a trait scale, socially desirable (positive) responses were confounded with the trait. In a semantic differential developed by Eiser and Osmon (1978), half the scales were anchored with positive labels, and half were anchored with negative labels. Positive anchored scales received significantly more extreme ratings than negatively anchored scales. The usefulness of balancing anchors appears to depend on what type of application the scales will have since these were all semantic differential scales.

Ory (1982) and Deaton, Glasnapp, and Poggio (1980) interpreted the balancing of items to mean that each item was either worded positively or worded negatively. Ory did not substantiate an influence in responses based on whether the item was positive or negative. Balancing items could not be supported in this context (students rating instructors). Barker and Ebel (1982) concluded that negatively worded items (on a true-false test) did not discriminate any better than positively worded items. Negatively worded items were designed to discriminate between the high and low achievers. The negatively worded items were found to be psychometrically more difficult to rate by the students than the positively worded items. However, they were not more discriminating.

Deaton, Glasnapp, and Poggio (1980) did find that item length and item direction main effects were statistically significant at the .05 level. When item length increased (more than 17 words), responses tended to be toward the center of the scale. When item length was short (less than 17 words), there was a tendency to respond toward the positive end of the scale resulting in higher mean responses. They concluded that items were ambiguous to the respondent when they were long and negatively worded. This appeared to influence respondents to rate these items toward the mid-range of the scale. Schuman and Presser (1981) included positive and negative statements in each item to construct a wholly balanced item instead of balancing items by placing only positive and only negative items in juxtaposition on a scale. They found no significant difference between their version of a balanced item and items that were not balanced (for national survey items). Balancing items did not appear to be useful when constructing the national survey items or in the construction of instructional rating scales. Personality trait measurements were influenced by balancing and length of questions.

Balancing items seems to be most influential when it is applied to the measurement of traits (Klockars, 1979; Deaton, Glasnapp, & Poggio, 1980). Ray (1982) substantiated that the traditional method of item reduction used in the construction of surveys would retain validity when items have been submitted to balancing.

CHAPTER V

DESIGN OF SCALE CATEGORIES

This chapter focuses on the design of scale categories. Several studies have been conducted to identify the best way to anchor a scale. Response alternatives selected have been varied such as: numbers, adjectives, adverbs, phrases, complete sentences, and descriptors of behavior. In selecting response alternatives, researchers must determine whether they wish to include the category generally known as the "Don't Know" category. This category would be useful for inclusion in a questionnaire for respondents who are not aware of the content of an item. The number of scale points to use is also an issue since there has never been consensus as to the optimal number of scale points. There have been recommendations for the use of a range of scale points all the way from 2 through 25. Obviously, this range includes scales that have an even number of scale points, as well as scales that have an odd number of scale points. When an odd number of scale points is selected, the labeling of the middle scale point position may cause difficulties for the researcher.

Apparently, the meaning of the middle scale point position has varied with respondents. The concept behind the middle position is that the midpoint indicates a halfway position on the bipolar scale. It is assumed that the middle position provides the respondents with a response alternative that allows them to rate an item as neutral. Yet, it is known that respondents will rate the middle position when they have no opinion at all. Because of this possibility, some researchers omit the middle response alternative altogether as a way to force respondents toward a polar position on the scale.

Labeling the middle response alternative has been of concern to researchers. It has been especially troublesome for those individuals tasked with developing behavioral scales. Since behavioral scales are built on large numbers of critical incidents, data reduction techniques are used to assign critical incidents to dimensions. Scaling the critical incidents generates more behavioral anchors toward the poles, leaving few at the midpoint. This has made it difficult to label the midpoint of the behavioral scales. The assignment of the midpoint response alternative has been ambiguous since different populations have divergent perceptions as to the meaning of the label. There have been suggestions to use terms such as "neutral" or "borderline."

There is no conclusive evidence to support the use of one specific number of scale points. It would be psychometrically acceptable to suggest a numerical range of acceptable scale points. A tentatively acceptable range might be between four and seven scale points. Five scale points are the most preferred and predominately used by researchers. The number of scale points is probably not what influences the reliability and the validity of a scale so much as the development of sound items. The same could be said for labeling a scale. Respondents seem to prefer scales with which they are most familiar, and are easy to use. This would be especially important for respondents that have lower levels of education.

Description of Response Alternatives

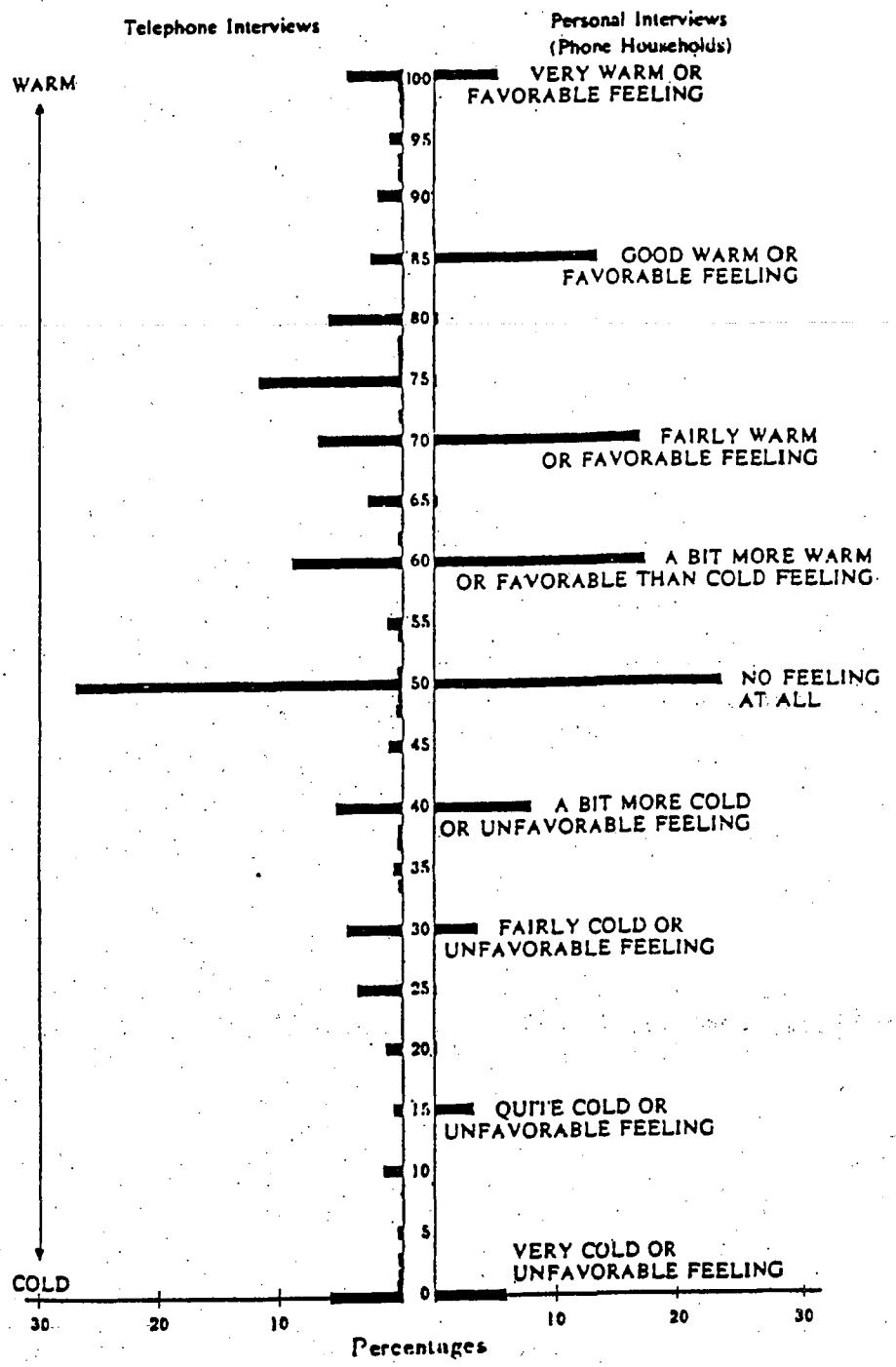
Points along the continuum of a scale have been identified (anchored) by numbers, adjectives, adverbs, description of behaviors, simple words, phrases, and complete sentences. Even the frequency and pattern of assigning anchors to the scale points has been varied. Some scales are completely anchored with an anchor at each scale point. Other scales have anchors only at the two endpoints of the scale. For example, the semantic differential has an anchor beyond each end of the scale which labels each bipolar direction.

Several studies have been conducted to discover the effects of different patterns and content of anchors on response distributions, reliability, etc. (Boote, 1981; Ivancevich, 1980; Borman & Dunnette, 1975; Reynolds & Jolly, 1980; Dolch, 1980; Menezes & Elbert, 1979; Mathews, Wright, Yudowitch, Geddie, & Palmer, 1978; Beltramini, 1982). Researchers have been interested in: the relative reliability of scales comprised of different anchoring, the cognitive structure used in responding to anchors, the abilities to define and differentiate among each anchor, and the raters preference for particular scales and anchors (Landy & Farr, 1980). The type of anchor selected may also be determined by how the questionnaire will be administered, the content area surveyed, and the population it is directed toward (Backstrom & Hurchur-Cesar, 1981; Groves, 1979).

Examples of Response Alternatives

Backstrom and Hurchur-Cesar (1981) developed anchors for items they considered sensitive or that they felt required complicated responses. They used cards having precoded responses printed on them for sensitive items, and also for a lengthy series of questions requiring complicated responses. Use of a response card with precoded alphabet letters for different categories allows the respondent to mention a specific category and tends to reduce respondent anxiety about revealing sensitive information. For a lengthy series of items with complicated response categories, they used a 7-point scale. The scale ranged from 1 (bad) through 7 (good). Each item was read to the respondents, and they were then requested to select a number from 1 through 7. The scale was printed on a card which the respondent held.

Groves (1979) reported on survey research conducted through personal interviews and over the telephone. The scale used consisted of a "political thermometer" anchored by degrees from 0 through 100 for items about Jimmy Carter. Groves indicated that labeling a point on a response card may facilitate its choice by a respondent. The following illustrated is Groves histogram of responses for the Carter feeling thermometer from telephone and personal interviews.



Groves (1979) found that telephone respondents tended to select numbers in the "political thermometer" that were divisible by 10. Respondents who were interviewed in person tended to cluster their responses around the labeled points on the response card.

Menezes and Elbert (1979) designed a questionnaire incorporating three scaling formats (semantic differential scale, Stapel scale, and Likert scale) to measure four itemized dimensions of store image. Illustrated here is their store image component measure for products using the three scales.

(A)
Semantic Differential Scale

	Extremely	Quite	Slight	Slight	Quite	Extremely	
Wide selection	:	:	:	:	:	:	Limited selection
Less known brands	:	:	:	:	:	:	Well known brands
High quality	:	:	:	:	:	:	Low quality

(B)
Stapel Scale

	+3	+3	+3
	+2	+2	+2
	+1	+1	+1
Wide selection		Less known brands	High quality
	−1	−1	−1
	−2	−2	−2
	−3	−3	−3

(C)
Likert Scale

	Strongly agree	Generally agree	Moderately agree	Moderately disagree	Generally disagree	Strongly disagree
Selection is wide						
Brands are less known						
Quality is high						

Reducing leniency was best accomplished by the Stapel scale while interrater reliability was highest for both semantic differential and Stapel scales. Each of the three scales have strengths in reducing rating errors. However, since they are not the same specific areas for reduction of errors, it is not possible to claim superiority for any one scale. Since each scale has a different format and is anchored differently, individual preference for format was solicited by Menezes and Elbert (1979).

Mathews, Wright, Yudowitch, Geddie, and Palmer (1978) conducted research on questionnaire response alternatives. The primary objective of the study was to establish the extent to which respondents' attitudes toward response alternatives were positive or negative on a bipolar scale of favorableness. The researchers thought that it would improve reliability if information were obtained on the favorableness of many candidate anchors. They developed lists of response alternatives which had descriptive terms delineating degrees of acceptability. These terms were presented to subjects to obtain norms regarding respondent perception of the response alternatives for: ambiguity, characteristics for degrees of acceptability, adequacy, and relative goodness. A secondary objective of the study was to take the normative data and construct sets of response alternative. The mean, standard deviation, and range of responses were used to select and space out the anchors, and thus reduce ambiguity of both input and output. They recommended the use of response alternatives which had smaller standard deviations. They concluded that response alternatives should be anchored at different points along the scale line so that they do not

overlap in the perception of the respondents. The term "borderline" was recommended as a response alternative for the midpoint in place of the term "neutral."

"Acceptability" descriptors are included below as examples of their research on response alternatives. The distribution of responses is described by mean, standard deviation, range, and number of subjects. Small standard deviations indicate consistency in perception by respondents for response alternatives, and would be more desirable as point anchors (Mathews, Wright, Yudowitch, Geddie, & Palmer, 1978).

RESULTS PERTAINING TO 'ACCEPTABILITY' DESCRIPTORS

Descriptor	Mean	SD	Range		No. of Subjects
			Min.	Max.	
Wholly acceptable	4.73	.56	3	5	51
Completely acceptable	4.69	.61	3	5	51
Fully acceptable	4.41	.87	2	5	51
Extremely acceptable	4.39	.72	3	5	51
Most acceptable	4.16	.92	2	5	51
Very very acceptable	4.16	.83	2	5	51
Highly acceptable	4.04	.63	3	5	50
Quite acceptable	3.22	.96	1	5	51
Largely acceptable	3.14	.99	1	5	51
Acceptable	2.39	1.46	0	5	51
Reasonably acceptable	2.29	.72	1	4	51
Moderately acceptable	2.28	.72	1	3	50
Pretty acceptable	2.00	1.13	-3	4	49
Rather acceptable	1.94	.82	0	4	49
Fairly acceptable	1.84	.92	0	4	50
Mildly acceptable	1.80	.95	-1	4	51
Mildly acceptable	1.69	.70	-1	4	51
Somewhat acceptable	1.46	1.24	-2	3	48
Barely acceptable	1.08	.52	-1	3	51
Slightly acceptable	1.04	.52	-1	2	51
Sort of acceptable	.94	.65	-1	2	50
Borderline	.00	.20	-1	1	50
Neutral	.00	.00	0	0	51
Marginal	-.12	.52	-2	1	50
Barely unacceptable	-1.10	.30	-2	-1	50
Slightly unacceptable	-1.26	.59	-4	-1	51
Somewhat unacceptable	-1.77	.67	-3	-1	51
Rather unacceptable	-2.02	.84	-4	0	50
Fairly unacceptable	-2.16	.88	-5	-1	50
Moderately unacceptable	-2.34	.68	-3	-1	50
Pretty unacceptable	-2.41	.66	-4	-1	51
Reasonably unacceptable	-2.44	.75	-4	-1	50
Unacceptable	-2.67	1.38	-5	0	51
Substantially unacceptable	-3.24	.90	-5	-1	51
Quite unacceptable	-3.39	1.07	-5	0	49

RESULTS PERTAINING TO 'ACCEPTABILITY' DESCRIPTORS (Cont.)

Descriptor	Mean	SD	Range		No. of Subjects
			Min.	Max.	
Largely unacceptable	-3.39	.82	-5	-1	51
Considerably unacceptable	-3.44	.78	-5	-2	50
Notably unacceptable	-3.50	1.04	-5	-1	50
Decidely unacceptable	-3.84	1.02	-5	-1	49
Highly unacceptable	-4.22	.58	-5	-3	50
Most unacceptable	-4.42	.72	-5	-2	50
Very very unacceptable	-4.49	.50	-5	-4	51
Exceptionally unacceptable	-4.54	.61	-5	-3	50
Extremely unacceptable	-4.69	.46	-5	-4	51
Completely unacceptable	-4.90	.36	-5	-3	50
Entirely unacceptable	-4.90	.36	-5	-3	50
Wholly unacceptable	-4.92	.27	-5	-4	51
Absolutely unacceptable	-4.92	.33	-5	-3	51
Totally unacceptable	-4.94	.24	-5	-4	51

Subsequent to this research, Dr. Charles Nystrom of the Army Research Institute, Fort Hood, suggested that an improved approach for the selection of response alternatives may be to use antonyms modified pairwise by the same pairs of adjectives or adverbs ("very satisfactory" and "very unsatisfactory;" "somewhat satisfactory" and "somewhat unsatisfactory", for example). Dr. Nystrom was able to obtain some (N = 30) judgments and opinions on what terms to use in rating scales containing 4, 5, 6, and 7 rating points.

As can be seen by the research, the study of anchors is extensive and includes many variations, such as alphabet letters, numbers, adjectives, adverbs, thermometers, etc., as well as many kinds of applications (U.S. Navy and Army officers and enlisted personnel, sales personnel, and marketing to households).

Comparisons of Response Alternatives

In a study previously mentioned (number of scale points), Boote (1981) performed market segmentation research with a mail survey to 600 households. Boote was concerned with scale points that were fully labeled or labeled at the extreme ends only. It was found that fully labeled scale points resulted in responses that were less skewed. The interpretation of this finding was that when scales are fully labeled, it promotes rejection of ratings which are closer to the extreme positive end of the scale. Landy and Farr (1980) reported research by Bendig (1952a, 1952b, & 1953) where the amount of scale anchoring increased the positive effect of the scales for performance appraisal.

Ivancevich (1980) performed research in the area of performance appraisal scales. He used subjects in sales from medium-sized organizations

on the east coast and mid-west. Ivancevich hypothesized that Behavioral Expectation Scales (BES) would exhibit less psychometric error than non-anchored scales or trait scales. Results indicated that BES was superior to nonanchored rating scales at the .05 level of significance for interrater reliability. Overall, psychometric superiority was not achieved through the use of the BES. Performance appraisals using behavioral anchors may not be worth the developmental effort. Ivancevich mentioned similar findings by Borman and Dunnette (1975) for subjects who were U.S. Navy personnel.

Market segmentation studies were conducted to evaluate three methods used to gather and evaluate value profiles with scales consisting of numerical ranks. Formats were developed for Likert ratings using 7-point scales and a paired condition using a minicomputer (Reynolds & Jolly, 1980). They found that the rank and Likert scales required less respondent time to complete than the paired-comparison method at the .001 level of significance. Interest in completing the scale items tended to decrease as the number of stimuli increased. Using Kendall's tau as a measure of test-retest reliability, the Likert scale was less reliable than rank order or paired-comparison scales. In another marketing study Menezes and Elbert (1979) evaluated three scaling formats (Likert scale, semantic differential scale, and Stapel scale) to measure store image. It was found that there were no overall differences among the three scale formats (each scale was anchored differently; see Menezes and Elbert for example of semantic differential, Stapel, and Likert scales).

Dolch (1980) compared semantic differential scales anchored by either numbers or adverbs, and concluded that there were high intercorrelations for both types of anchors. There appeared to be no difference between anchors. However, when the semantic space was factor analyzed, it appeared that the adverbial anchors had different meanings for different respondents. Apparently, the two scales were not measuring meaning in the same way. In research performed by Beltramini (1982), the following scales were compared: unipolar versus bipolar, 5 through 10 response alternatives, and horizontal versus vertical physical format. Some of these scales were comprised of verbal anchors and some consisted of numerical anchors. Beltramini (1982) found that none of the main or interaction effects were significant at the .05 level.

Inconsistency of results for application, scale construction, scale format, and scale anchoring suggests that perhaps the research would produce more useful results if scale item investigations were pursued in lieu of response alternatives. The assumption is that good scale item construction will be followed by the selection of anchors that are definitive so that respondents will not attribute the same meaning to more than one scale point along the continuum. Mathews, Wright, Yudowitch, Geddie, and Palmer (1978) proposed that scale anchors should occupy narrow bands along the scale continuum so that they do not overlap. This is why they only selected anchors which had a standard deviation of 1.00 or smaller.

Conclusions Regarding Response Alternatives

There are any number of ways a scale can be anchored (alphabet letters, numbers, political thermometer 0 to 100 degrees, verbal anchors, and

behavioral anchors). Marketing studies comparing different scales and different anchors (Reynolds & Jolly, 1980; Menezes & Elbert, 1979) were not able to find overall differences across scaling formats. Scaling developed for performance appraisal (Ivancevich, 1980; Borman & Dunnette 1975; Landy & Farr, 1980) comparing different scale formats and different anchors indicated that no one format was able to claim psychometric superiority over another. It was suggested by Landy and Farr that the best type and number of anchors selected would probably depend on the adequacy of the scale dimensions.

There has also been an inconsistency for the reliability and validity of an instrument and the preference of respondents for instrument usage. For example, Menezes and Elbert (1979) determined that each scale has its own strengths and weaknesses, and that no one scale could be claimed as being more robust than another (Likert, semantic differential, Stapel). They questioned which scale would be of most use in measuring retail images. Respondents in this study ranked the Likert scale as most preferred followed by the semantic differential, and lastly the Stapel scale. They suggested that the easiest formats be selected for less educated subjects. For ease of scale construction, the Stapel scale ranks first since it alleviates the problem of selecting antonyms or constructing Likert-type statements.

There is some evidence (Boote, 1981; and Bendig, 1952a, 1952b, 1953) that anchoring scales is useful in obtaining superior psychometric results. However, this area of investigation has received little replication for the number of scale points anchored, and there has been great inconsistency in results to support any one type of anchoring system versus another. If anchors are selected independent of any item and measured for bands along the scale dimension, there is the potential that anchor linkage to the item would modify the standard deviation of each anchor.

Beltramini (1982) and Dolch (1980) anchored scales verbally and with numbers. In both cases, no one scale was psychometrically superior to another. Variations in the anchors did not seem to affect the item's ability to discriminate. Dolch determined that the semantic space was different for adverb versus numerical anchors. The developmental procedures used in selecting the items may be of greater importance than the anchoring since similar results have been obtained using different anchors. The determination of which type of anchor to use should also be contingent on the questionnaire application (survey use, appraisal, description of respondents, etc.).

5.2 "DON'T KNOW" CATEGORY

Description of the "Don't Know" Category

Some respondents are known for their tendency to withhold an opinion. They have a tendency to prefer to mark the category "don't know" when it is an option on questionnaire forms. Withholding an opinion could mean that the respondent is not aware of the content in the questionnaire item and has no knowledge of the content area. Another interpretation when selecting the "don't know" category is that the respondents refuse to express their opinion (Backstrom & Hurchur, 1981). Many attempts have been made to determine the personality trait profile of respondents who have the tendency to select the "don't know" category (Innes, 1977; Biggs, 1970; Schuman & Presser, 1981). However, results of research have been inconsistent in the identification of a specific personality trait or a demographic attribute, such as age, sex, education, etc.

It has been determined that a certain strata of respondents will provide a substantive response to a standard version of a questionnaire form (that does not have a "don't know" category). Yet, they will include "don't know" when they are provided the opportunity. These same subjects will indicate a "don't know" response when it is included in their selection choice on the form. To measure the "don't know" response, Schuman and Presser (1981) developed "filtered" questions along with standard questions. The filtered questions have an option for the "don't know" category where standard questions do not. It is possible for subjects to volunteer a "don't know" response on the standard form.

Examples of the "Don't Know" Category

Schuman and Presser (1981) established "don't know" filter items on various surveys to identify what type of respondent would select an opinion on one questionnaire form (without a "don't know" category) and then mark a "don't know" on surveys that include that option. Examples of their filter and standard questions are provided. These questions were previously incorporated into surveys from the National Opinion Research Center (NORC) and the Survey Research Center (SRC). Included along with the questions are the marginals. (Marginals are the percentage of responses to each response alternative for each item in a questionnaire.)

Schuman and Presser "Don't Know" Filter Experiments

1. Courts (NORC-74)

Standard Form		Filtered Form	
"In general, do you think the courts in this area deal too harshly or not harshly enough with criminals?"		"In general, do you think the courts in this area deal too harshly or not harshly enough with criminals, or don't you have enough information about the courts to say?"	
Too harshly	5.6%	Too harshly	4.6%
Not harshly enough	77.8%	Not harshly enough	60.3%
About right (volunteered)	9.7%	About right (volunteered)	6.1%
Don't know (volunteered)	6.8%	Not enough information to say	29.0%
	(N=745)		(N=723)

2. Government (SRC-76 February)

Standard Form		Filtered Form	
"Some people are afraid the government in Washington is getting too powerful for the good of the country and the individual person. Others feel that the government in Washington is not getting too strong. What is your feeling, do you think the government is getting too powerful or do you think the government is not getting too strong?"		"Some people are afraid the government in Washington is getting too powerful for the good of the country and the individual person. Others feel that the government in Washington is not getting too strong. Have you been interested enough in this to favor one side over the other? (If yes) What is your feeling, do you think the government is getting too powerful or do you think the government is not getting too strong?"	
Too powerful	55.0%	Too powerful	45.0%
Not too strong	35.1%	Not too strong	21.6%
Don't know (volunteered)	10.0%	Not interested enough	33.3%
	(N=613)		(N=606)

3. Communist Book (SRC-77 February)

Standard Form		Filtered Form	
"This next question is about a man who admits he is a communist. Suppose he wrote a book which is in your public library. Somebody in your community suggests the book should be removed from the library. Would you favor removing the book or oppose removing the book?"		"This next question is about a man who admits he is a communist. Suppose he wrote a book which is in your public library. Somebody in your community suggests the book should be removed from the library. Would you favor removing the book, oppose removing the book, or do you not have an opinion on that?"	
Favor removing	29.1%	Favor removing	17.2%
Oppose removing	67.9%	Oppose removing	56.6%
Don't know (volunteered)	3.0%	No opinion	26.2%
	(N=563)		(N=533)

Comparisons of "Don't Know" Categories

Research performed by Innes (1977) focused on the extremity of the response set and the "don't know" response for questionnaire items. The subjects were male students enrolled in a technical college. Results of this research indicated that there appeared to be a "don't know" response set which was correlated with measures of originality and divergent thinking. Innes postulated that more creative individuals would be prone to reserve judgment and select a "don't know" category since they would be able to accept ambiguity.

Schuman and Presser (1981) conducted research to identify respondents who would give a substantive response to a standard item that did not have a "don't know" category. These respondents would shift over to a "don't know" response when it was offered. These respondents were termed "floaters." Nineteen experiments were conducted using items from NORC and SRC. They found that, with the rewording of questions to include a "don't know" response, it is possible to shift responses for more than a fifth of the subjects on a consistent basis. The distribution of the substantive responses was not significantly different for the standard items compared to the filtered items when the "don't know" responses were eliminated for 14 of the 19 experiments. They concluded that filtered items do not usually affect research results.

Subjects who will shift over to a "don't know" response have not been identified by trait or traits (Schuman & Presser, 1981) as they were with Innes (1977). There is a correlation between "don't know" responses and low education, as well as lack of topic information. Ambivalence may also be a variable influencing the "don't know" response. These variables are not predictive across studies. They were not able to identify any special trait, traits, or group. Researchers are not able to predict who would make a shift into a "don't know" category. The content of each item may contribute to "don't know" responses associated with the familiarity of issues.

Conclusions Regarding the "Don't Know" Category

Fourteen experiments out of 19 were not able to identify a trait, traits, or a group that shifts their responses over to "don't know" (Schuman & Presser, 1981). One experiment (Innes, 1977) found a trait related to the "don't know" response. It is not possible to predict in advance what individual or group of individuals is going to make a "don't know" response.

The other five experiments obtained significant differences between the standard version (did not include "don't know" category) and the filtered version (included "don't know" category) ranging between .05 and .001 levels of significant. Schuman and Presser (1981) concluded that including a "don't know" category can at times (on a limited basis) alter the dispersion of opinion data. However, the "don't know" category typically does not alter opinion; and when it does, its effect is usually small. They determined that a low level of education was not correlated with respondent selection of the "don't know" category in most situations. The researchers were not able to identify these "don't know" respondents by personality or

social characteristics. It appears as though the content of the survey item may influence the selection of a "don't know" response for items dealing with obscure issues. For this type of item, there is a correlation with respondents identified as having a low level of education (0 to 11 years of school).

Apparently, knowledge of the "don't know" response set does not significantly influence the response distributions when the "don't know" responses are eliminated from the questionnaire (in most cases). The actual content of the item may be determining the likelihood of a "don't know" response for items which have unfamiliar content to the respondents. There appears to be no special set of individuals who will shift (when given the opportunity) over to a "don't know" response. There is a relationship between low education and selection of the "don't know" response for obscure issues. The same holds true for individuals with a weak opinion on a topic or a lack of information about a topic.

5.3 NUMBER OF SCALE POINTS

Description of Number of Scale Points

In questionnaire construction, researchers have investigated the utility of having a scale with a greater or smaller number of scale points. Selection of scale point number ultimately hinges on how many scale points are best to achieve the researcher's objectives. Over the years, there have been diverse recommendations on the proper number of scale points or categories to use in questionnaire construction. Comrey and Montag (1982) reported research by Symonds (1924), Nunnally (1967), Garner (1960), and Guilford (1954) which indicated that reliability was optimum for scale points of 7, 11, 20, and 25. More recent research has proposed the use of a range of scale points between 2 and 10 (Schutz & Rucker, 1975; Beltramini, 1982).

Studies for determining scale point number have focused on the type of application. For example, Guion (1979) suggested using a small number of scale points for personnel testing to measure representation of real world situations. How the scale points are anchored has also been investigated. Boote (1981) found that fully-labeled scale points achieved greater reliability than scale points where only the extremes were anchored. The selection of number of scale points is dependent on the type of application, the anchoring format, and the quality or ability of the scale anchors to differentiate among conditions.

Examples of Number of Scale Points

Research performed in the areas of human factors engineering, advertising, and marketing research provides examples of scales with different numbers of scale points.

Illustrations of items designed for a 2-point scale and a 5-point scale are provided for the area of human factors engineering, vehicle maintenance, amphibious operation (Krohn, 1984). The 2-point and 5-point scales include an additional category for "not applicable" or "not observed." Following is a portion of an interview outline for amphibious operation developed by Krohn:

> "I will name equipment from the LAVM/RV that you may have used to perform amphibious operations. Please answer Yes or No to indicate whether or not you experienced any difficulties using the equipment. I would also appreciate your comments concerning the difficulties. If you have no experience using the equipment, then check the Not Applicable column."

Equipment	Yes	No	NA	Comment
Propellers	___	___	___	_____
Rudders	___	___	___	_____
Rudder Controls	___	___	___	_____

The 5-point scale developed by Krohn (1984) used a variation of the "Nystrom Number Scale" (reported in Questionnaire Construction Manual for Operational Tests and Evaluation (Church, 1983) and developed by Dr. Charles Nystrom of the Army Research Institute, Fort Hood, Texas. An example of the original Nystrom Number Scale (Church, 1983) is followed by the Krohn (1984) version:

Nystrom Number Scale

15.	"Wires & seals on external fire extinguisher handles.	EASY	+2	+1	0	-1	-2	DIFFICULT
		( ) Not Checked						
16.	That Fire Suppression switch is in AUTO.	EASY	+2	+1	0	-1	-2	DIFFICULT
		( ) Not Checked						
17.	Wire & lead seal on internal fire extinguisher.	EASY	+2	+1	0	-1	-2	DIFFICULT
		( ) Not Checked						
18.	For open or missing front hull drain plug.	EASY	+2	+1	0	-1	-2	DIFFICULT
		( ) Not Checked						
19.	For open or missing rear hull drain plug."	EASY	+2	+1	0	-1	-2	DIFFICULT
		( ) Not Checked						

Maintenance Vehicle Questionnaire by Krohn

Ease of Use Rating Scale

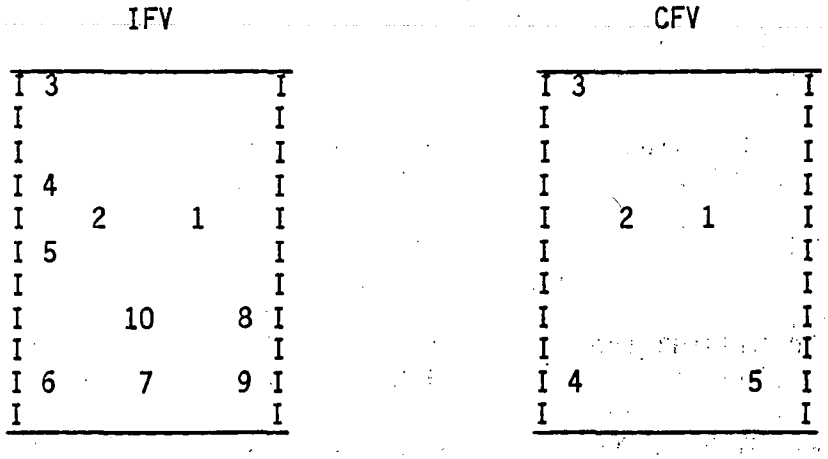
5	4	3	2	1	N
Very Easy	Easy	Borderline	Difficult	Very Difficult	Not Applicable or Observed

How easily can you:

1.	"Gain access to the vehicle's batteries?	5	4	3	2	1	N
2.	Check battery and fluid levels?	5	4	3	2	1	N
3.	Check tightness of battery cables?"	5	4	3	2	1	N

The Nystrom Number Scale includes directions for the respondents. An illustration for respondent directions is presented here for the operation under usual conditions, Bradley Fighting Vehicle Test:

> "Please show your duty location in the BFV by drawing a circle around your seat location number in the appropriate IFV or CFV diagram below. If you are the track commander, gunner, or driver, circle the 1, 2, or 3, respectively."



> "Questions 10 through 86 all identify tasks performed when operating under usual conditions. For each item, please rate how easy-difficult it is to perform the task named. Circle just one of the numbers (+2, +1, 0, -1, -2) for each question, or check this ( ) if you have not performed the task."

The Nystrom Number Scale illustrated uses five scale points, +2 through -2. It would be possible to construct the scale using 4, 5, 6, 7, 8, or 9 numbers between the anchor words. Following are two examples of the Nystrom Number Scale with varying five and seven scale points:

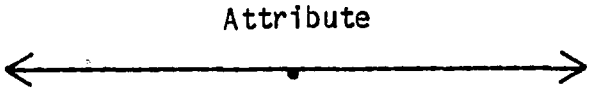
EASY +2 +1 0 -1 -2 DIFFICULT
( ) Not performed

INEFFECTIVE | EFFECTIVE
-3 -2 -1 0 +1 +2 +3

Beltramini (1982) compared unipolar versus bipolar, number of scale intervals (5 through 10), and horizontal versus vertical scale formats in measuring scalability to discriminate between two advertisements for a

national fast-food restaurant. Three illustrations of the basic rating scales used by Beltramini (1982) are provided:

Basic Rating Scales



Bipolar, 5-interval, Horizontal Scale

good - - - - - bad

Unipolar, 10-interval, Vertical Scale

Humorousness

+5 ___
+4 ___
+3 ___
+2 ___
+1 ___
-1 ___
-2 ___
-3 ___
-4 ___
-5 ___

Four scales were developed by Schutz and Rucker (1975). Each scale was anchored at the extreme ends by the terms "appropriate" and "inappropriate." The numbers of scale points compared were 2, 3, 6, and 7 for a food-use questionnaire. They were interested in determining how respondents felt about the appropriateness of different foods in a number of situations. Respondents were presented with a grid that listed food across one side and various situations across the other side. For each food-use combination on the grid, respondents selected a number that represented the appropriateness of the combination. In the example presented below, if a respondent felt it was appropriate to eat jello while watching TV, he or she would place a "1" in the top left-hand cell on the grid (1 indicating "appropriate" and 7 indicating "inappropriate").

> "Please fill in the grid working down the columns. Preliminary research indicates that filling in each column before going on to the next item is faster than working across the rows. You may not be familiar with some of the foods or have engaged in some of the uses or food-use combinations. Even if that is the case, for each food-use combination, please give us your opinion of how appropriate it is to use this food in this situation. Do not leave any cells blank. Since we are interested in your opinion regarding appropriateness, we would appreciate it if you would

ratings between the first and second mailing. A sample of Boote's questionnaire items along with the correlation coefficients for respondents' scale ratings is provided for scale formats having 5 and 7 points:

	Scale formats			
	All pts. labeled		Only extreme pts. labeled	
	5 pts.	7 pts.	5 pts.	7 pts.
"Having a familiar routine for getting things done.	.586	.658	.659	.710
Doing things the best way even if it takes longer	.597	.635	.552	.391
Getting away from my home occasionally to enjoy my leisure time	.632	.697	.423	.540
No matter what I buy, to have only the best that I can afford	.567	.535	.396	.532
To have clothes which fit properly	.658	.526	.122	.448
Having lots of different models to choose from when I buy an appliance" . . .	.594	.573	.439	.580

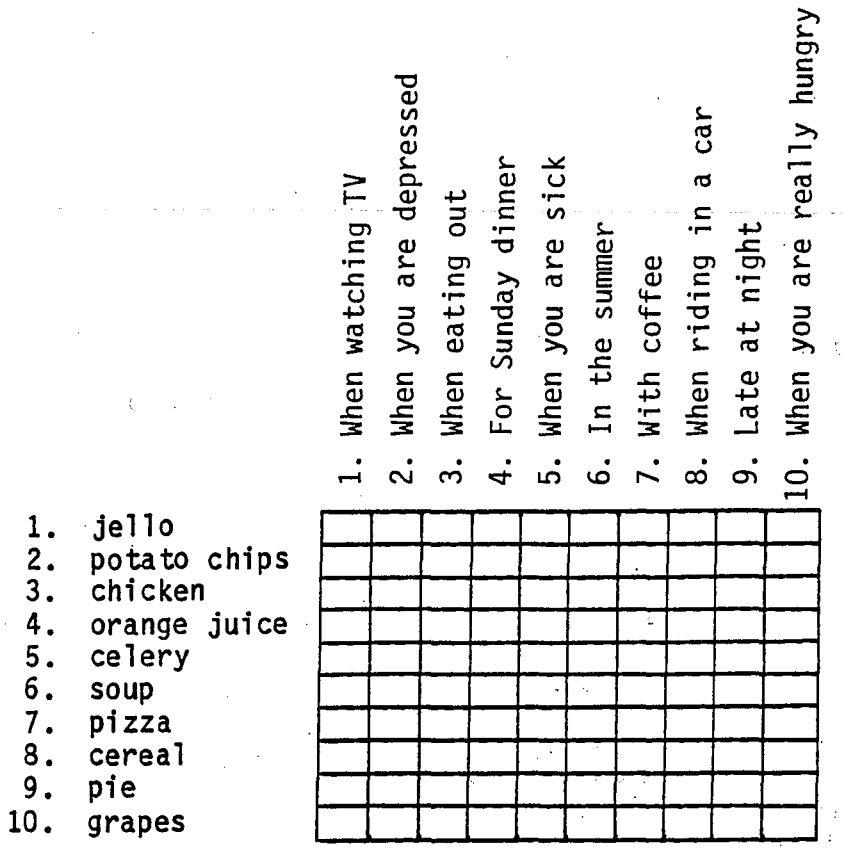
Comparisons of Number of Scale Points

Studies consistently vary formats for number of scale points compared, types of anchors used/or not used, and the actual areas of application. Conflicting evidence indicates that in some instances, the number of categories does not affect responses to a scale (Schutz & Rucker, 1975; & Beltramini, 1982). Other investigations have yielded definite preferences for number of scale points (McKelvie, 1978).

Beltramini (1982) analyzed physical format using 24 cells and 1,296 subjects in the following format variations: 5 - 10 scale points, unipolar versus bipolar, and horizontal versus vertical questionnaire formats. He assessed the ability of the format variations to discriminate between two advertisements used by a national fast-food restaurant. No interaction effects were significant at the .05 level. Differences in number of scale points (5 - 10), polarity, or physical format (horizontal/vertical) alone or in interaction did not affect the scale's ability to discriminate between advertisements. These results indicate that variations in scale format are not the critical issue in scale development. The manipulation of physical format by number of scale points (2 - 7) was investigated by Schutz and Rucker (1975). These scales were anchored at the extremes on a food-use questionaire. They came to a similar conclusion to that of Beltramini (1982). Correlation coefficients were .98 or higher for all scales which suggests that the number of scale points does not change the cognitive approach by subjects in rating items.

In a marketing study by Boote (1981), four scale formats were compared: 5-point scale labeled at each scale point; 5-point scale labeled

not compare your responses with those of other people until after you have completed the grid."



People's value orientations were measured by Boote (1981) using 5-point and 7-point scales where either all scale points were anchored or only the extremes were anchored. Boote developed four different formats and evaluated them for test-retest reliability. Format 1 consisted of five scale points which were labeled "extremely important," "very important," "somewhat important," "slightly important," and "not at all important." Format 2 consisted of five scale points which were anchored only at the endpoints by "extremely important" and "not important at all." Format 3 consisted of seven scale points which were labeled "extremely important," "very important," "quite important," "somewhat important," "moderately important," "slightly important," and "not at all important." Format 4 consisted of seven scale points which were anchored only at the endpoints by "extremely important" and "not at all important." For the test-retest condition, the questionnaires were mailed out. After six weeks, a second mailing of the same questionnaire was sent out to the same individuals who responded to the first questionnaire. The correlation coefficients found in a reduced version of the example shown next are for respondents' scale

only at the endpoints; 7-point scale labeled at each scale point; and 7-point scale labeled only at the endpoints. Boote examined differences in reliability attributed to differences in anchoring scale points and number of scale points. It was determined that scales fully labeled yielded less skewed response distributions than scales labeled only at the endpoints. Five-point scales were superior to 7-point scales for these particular marketing studies.

McKelvie (1978) concurred with Boote in that a recommendation was made for the use of five or six scale points, but no greater or lesser number than five or six. It was felt that a greater number of scale points would have no psychometric advantage, and that a smaller number of scale points would threaten the discriminative power and validity of the instrument. McKelvie's research was conducted using instruments measuring opinions as well as psychophysical stimuli (tones).

Boote (1981) and Mckelvie (1978) have divergent recommendations when it comes to labeling the scale points. Boote's findings indicated that it was best to label each scale point, and McKelvie's findings indicated that it made little difference regarding the reliability and validity of an instrument whether verbal labels were used. Their samples were composed of different populations (students versus respondents from households). The applications were quite divergent (marketing psychographic segmentation, public opinion, and psychophysical measurements).

Simulation of test scores by Lissitz and Green (1975) using a multivariate normal generator with different numbers of scale points (2, 3, 5, 7, 9, and 14) resulted in increases in the standard deviation as covarience decreased, and decreases in the standard deviation as the number of scale points increased. They found a leveling off in the increase of reliability after five scale points. They reject 7-point scales as an optimum number and support the use of 5-point scales.

In the comparison of personality item formats for a 2-choice or 7-choice response format, Comrey and Montag (1982) concluded that the 7-choice response (7 scale points) allowed for finer discriminations by subjects using a personality inventory. In this study, five scale points were not included as one of the format variations.

In selecting the number of scale points to use in a study, the selection will depend on the area of application. There is a trend toward the use of 5-point scales. Five-point scales were recommended for the development of tests (Lissitz & Green, 1975), marketing surveys (Boote, 1981), and measuring psychophysical stimuli (McKelvie, 1978).

Conclusions Regarding Number of Scale Points

The number of scale points selected will depend on the research design, the area of application, and the types of anchors used. However, the developmental procedures used in the design of items probably has more weight than the physical format which would be represented by the number of scale items and types of anchors (Beltramini, 1982; Schutz & Rucker, 1975).

There is some psychometric support for the selection of five scale points as an optimum number across areas of application (Boote, 1981; McKelvie, 1978; Lissitz & Green, 1975). Even so, because of conflicting evidence from studies recommending seven scale points (Comrey & Montag, 1982), a range of five or six scale points (McKelvie, 1978), or greater ranges of scale points (all the way from 2 through 10) (Schutz & Rucker, 1975; Beltramini, 1982), it is not possible to recommend with certainty a specific number of scale points. There is flexibility within the selection process.

There is no conclusive evidence to support which is the best way to anchor the scales once the number of scale points has been identified. McKelvie (1978) found no significant effect for anchoring, but Boote (1981) found that fully-labeled scale points achieved higher reliablity than anchoring only the extreme endpoints of a scale. As with the number of scale points, there is flexibility in selecting the scale anchors since research trends have not been able to identify optimal response alternatives. There has been a shift in research so that a greater emphasis has been placed on developmental procedures for items and anchors, training of raters, and cognitive approaches to rating by subjects.

5.4 MIDDLE SCALE POINT POSITION

Description of Middle Scale Point Position

The middle position on a bipolar scale can be used to provide respondents the opportunity to rate a system or a thing as between "satisfactory" and "unsatisfactory," between "adequate" and "inadequate," between "effective" and "ineffective," etc. In these instances, the middle response option corresponds to the zero point on an algebraic scale. It's like a point between two intervals, although one may also view it and treat it as an interval between two other intervals.

It has been questioned whether to use a midpoint in scale construction or whether it would be better to construct scales with an even number of scale points. Presser and Schuman (1980) found that when a middle position is offered on a scale, there is a shift of respondent ratings into that midpoint by up to 10-20% or larger. In addition, there is only a slight decrease in the "don't know" category when a middle alternative is offered. The shift to the midpoint apparently comes from the polar positions.

In situations where researchers have elected to use a middle alternative, anchoring the midpoint has been an issue. Ideal scale anchors are located along the scale with meanings that produce response distributions that they do not overlap so that respondents don't become confused and attribute the same meaning to more than one scale point (Mathews, Wright, Yudowitch, Geddie, & Palmer, 1978).

Some researchers intentionally omit the middle alternative as a way of forcing respondents toward a polar position on the scale. It is possible in this context to have the structure of the scale shift the respondent's selection of a response alternative (Presser & Schuman, 1980).

Examples of Middle Scale Point Position

Dollard, Dixon, and McCann (1980) designed a student questionnaire for the evaluation of the Automated Shipboard Instruction and Management System that was used aboard the U.S.S. Gridley. The questionnaire combined checklists and items with response alternatives that omitted the middle position. Following are illustrations of questions which omitted the middle alternative. These questions offered responses that included "yes," "no", and "?." The "?" was to indicate "don't know" or "non-applicable."

	Yes	No	?
"Did your divisional DCPO or PQS qualifying petty officer ever help you with your CII course when you needed assistance?"	___	___	___
"Do you intend to reenlist when your present enlistment expires?"	___	___	___

Presser and Schuman (1980) designed a series of experiments to measure the effects of the middle position in attitude surveys. They used two forms for each item where one form had a middle position and the other form did not. The items they used were selected from the Gallup Survey, Institute for Social Research, National Opinion Research Center, and Survey Research Center. Following are modified versions of items that do and do not include a middle position.

Do you feel that the state government has too much or too little control over local law enforcement training?

1. Too Much 5. Too Little 3. (If Volunteered) Right Amount

Do you feel that the state government has too much, too little, or the right amount of control over local law enforcement training?

1. Too Much 5. Too Little 3. Right Amount

Mathews, Wright, Yudowitch, Geddie, and Palmer (1978) developed a list of scale anchors. These included midpoint anchors which did not overlap or were minimally overlapping along the scale continuum. The criteria that they established for anchor selection was that no anchor was selected if the standard deviation was 1.00 or greater. Anchors having the largest means were selected for the positive extreme end of the scale. The other anchors were selected in a descending order. Anchors were to be at least one standard deviation apart. Following are three sets of anchors they identified that have minimally overlapping descriptors for acceptability, adequacy, and relative goodness:

Descriptor	Mean	SD
Wholly acceptable	4.73	.56
Highly acceptable	4.04	.63
Reasonably acceptable	2.29	.72
Barely acceptable	1.08	.52
Neutral	.00	.00
Barely unacceptable	-1.00	.30
Somewhat unacceptable	-1.77	.67
Substantially unacceptable	-3.24	.90
Highly unacceptable	-4.22	.58
Completely unacceptable	-4.90	.36

Descriptor	Mean	SD
Totally adequate	4.62	.85
Very adequate	3.42	.85
Reasonably adequate	2.41	.77
Mildly adequate	1.57	.67
Barely adequate	.63	.93
Barely inadequate	-1.16	.64
Somewhat inadequate	-1.88	.73
Considerably inadequate	-3.60	.68
Very very inadequate	-4.46	.54

Descriptor	Mean	SD
Best of all	4.90	.51
Extremely better	3.92	.88
Moderately better	2.26	.74
Slightly better	1.16	.78
Alike	.22	.85
Barely worse	-1.04	.82
Somewhat worse	-2.08	.86
Conspicuously worse	-3.28	.89
Absolutely worse	-4.43	.82

Comparisons of Middle Scale Point Position

Presser and Schuman (1980) hypothesized that respondents using a form that did not have a midpoint would respond in a similar way to respondents that did have a midpoint on their questionnaire. If this were true, the frequency counts (and percentages for each scale point) would be similar for both forms with the exception of the middle response category. Ten experiments were conducted to test this hypothesis. A significance level was not reached for any of the 10 experiments. This indicated that the percentages for each category were similar on both forms whether the midpoint is excluded or included.

The decline in polar positions accounts for the shift in response when a middle alternative is offered. Presser and Schuman (1980) and Schuman and Presser (1981) indicated that the level of intensity of opinion is a factor in determining whether respondents are affected by the form structure for a midpoint or lack of a midpoint. They have found that information on content area and level of education appear to be unrelated to the form effect. More intense respondents (individuals with a strong opinion on the topic) exhibit less form effect than respondents who are less intense or have no opinion. Presser and Schuman (1980) and Schuman and Presser (1981) suggested that more intense respondents would be less influenced to rate the midpoint. Some high intensity subjects did exhibit a response shift for scales with a midpoint.

Presser and Schuman (1980) and Schuman and Presser (1981) noted that middle alternative anchors are generally used for surveys. They recommended that examination of anchors for the middle alternative would be useful in conceptually defining populations. This type of research was performed by Gividen (1973) and Mathews, Wright, Yudowitch, Geddie, and Palmer (1978). Their subjects were Army officers and enlisted men. Mathews, et al. investigated verbal anchors for scale value on a bipolar scale of favorableness (from positive to negative). (See identified descriptors for acceptability, adequacy, and relative goodness for verbal anchors provided as examples in this section.)

Research results for scale midpoints were obtained with means and standard deviations at zero for the three lists of scale anchors developed that had a midpoint termed "neutral." Mathews, Wright, Yudowitch, Geddie, and Palmer (1978) cited previous research by Gividen (1973) where Army test officers were not totally clear regarding the meaning of the term "neutral" as a midpoint anchor. Some respondents thought it meant indifferent, having no opinion. There were respondents who thought it meant the value in the middle of the scale. Others were aware that it could mean either and didn't know which meaning was intended. Because of the ambiguity surrounding this term, Gividen recommended the term "borderline" as a midpoint anchor. The term "borderline" was coined by Dr. Charles Nystrom of the Army Research Institute, Fort Hood, Texas.

The design of rigorous scales requires examination of the item variability for verbal anchors since there can be large variances among subjects in their assignment of values to anchors. Scale values obtained by Mathews, Wright, Yudowitch, Geddie, and Palmer (1978) cannot be directly generalized over to other questionnaires.

Conclusions Regarding Middle Scale Point Position

Researchers are simultaneously confronted with two issues in the construction of questionnaires as it relates to the scale midpoint. The first issue is whether they wish to include a midpoint in their scale. The second issue is, if they do, then how should the midpoint be anchored? Investigations in the area have not been abundant although previous research does provide some guidelines.

It has been common practice in the construction of questionnaires to eliminate the middle alternative in an effort to force respondents toward one or the other poles on a bipolar scale. In these situations, there is the possibility that the format may be assisting in structuring the respondent's decision-making. This may be especially true for attitude questionnaires when respondents have weak opinions, or have no opinion regarding the content of the question.

Presser and Schuman (1980) found that response distributions that include the middle response alternative look about the same as the distributions without the middle response alternative. The decision regarding the inclusion or exclusion of the middle category in the design of a scale should depend on the type of information the researcher is interested in retrieving from the subjects. When the researcher seeks a highly refined/ precise description of the response distribution, the inclusion of a middle

alternative would be useful. Of course, use of additional polar categories would also support this objective. It may be better to exclude the middle alternative when subjects have a weak opinion on the topic and/or it is of importance to elicit the direction of the opinion/ attitude. Forcing respondents toward one of the poles may be viewed as a trade-off. Those respondents with weak opinions tend to select the middle alternative in large proportions. When the middle alternative is deleted from the response alternatives, they are forced toward a polar position. Some respondents may indicate that they were not allowed to accurately state their opinion.

As to the identification and selection of the midpoint anchor, it is not unrelated to the other anchors used on the scale. It is not possible to identify a midpoint anchor without also determining the content and form of the other scale anchors as well. In the research performed by Gividen (1973) and Mathews, Wright, Yudowitch, Geddie, and Palmer (1978), they identified the midpoint anchors "borderline" and "neutral." Mathews, et al. indicated the preference for the usage of "borderline" based on results obtained by Gividen.

Different populations are going to have different means and variances in their attitudes toward various scale anchors whether they be midpoints or located at the extreme ends of the scales. Recent studies have indicated that a sound scale is predicated on the developmental procedures used in the construction of the items more than on the format or type of anchor used.

CHAPTER VI

INTERVIEWER AND RESPONDENT CHARACTERISTICS

The effect of the interviewer on respondent ratings is examined in this chapter. The impact of demographic characteristics on response distributions is also reviewed. It has not proven feasible to identify any one questionnaire format over another. There have been suggestions by questionnaire construction experts that other characteristics related to the respondent may be more potent and reliable in the design of questionnaires. Investigators have been trying to enhance the psychometric quality of ratings by adapting the rating format to the cognitive structure of the respondent. This approach takes into account the respondent compatibility with the demands of the rating format. This form of questionnaire construction has been termed cognitive complexity when applied to behavioral scales.

Four demographic characteristics are broken out into individual sections (education, ethnic background, age, and gender) in Chapter VI. It is common knowledge that these variables frequently interact with each other in empirical investigations. There is some evidence that response patterns are influenced by the education of respondents for high levels of education and for low levels of education. High educational level has been defined as completing at least some college. Low educational level has been defined as not completing high school. The research indicates that individuals with a low educational level may be the most influenced by item wording or survey format. Individuals with a low level of education may also be prone to survey nonresponse when education is interrelated with other variables.

Ethnic background of respondents has been examined for its effect on rating patterns. For surveys which use interviewers, the influence of the ethnic background of the interviewer has been investigated. This research tends to indicate that nonracial items appear to be immune to interviewer effects for ethnicity. Performance ratings have also supported these findings where no significant differences were found between black and white raters. There have been exceptions to this finding in the area of self-assessment.

Item rating is sometimes influenced by age, education, and item content. This may be a phenomenon of opinion questionnaires. These questionnaires could easily elicit different responses, depending on the perceptions of different age and gender groups. When item responses are influenced by the age of the respondent, it most often relates to item nonresponse or survey nonresponse by older subjects. As with other demographic characteristics, age and education usually interact with other demographic characteristics.

Description of Interviewing

The conduct of a survey through interviews is sometimes made in preference to a mail survey or to a paper/pencil questionnaire administration. There are several situations which would support the choice of interviews. For example, it is well known that telephone interviews and face-to-face interviews have a higher response rate than mail surveys (Orlich, 1978; Shosteck & Fairweather, 1979). In situations where a high survey response rate is critical, the interview would be a primary vehicle for achieving that purpose. When survey results are required within a short period of time, it is possible to use telephone interviews. The Air Force has been known to use telephone surveys where the results were reported within 48 hours (Chun, Fields, & Friedman, 1975). Air Force personnel have also used interviews for survey data collection. Pilots served as interviewers for respondents who were test pilots. This approach was suggested to reduce error. It was thought that the questionnaire might not fully reflect the pilot's experience or opinion. The interviewer would be able to probe the test pilots to determine the in-depth meaning of their responses (Church, 1983).

One of the drawbacks to the use of interviews has been the increased cost compared to surveys that do not require interviewers (Orlich, 1978). There is the cost of training the interviewers, the cost of sending the interviewer to the face-to-face interview site, and the time involved for each interview. Shosteck and Fairweather (1979) compared the cost of mail surveys with that of surveys using face-to-face interviews. Mail surveys were $24 per respondent, and interviews were $63 per respondent (these figures do not include administrative costs).

Researchers interested in obtaining accurate data from their interviews generally ask multiple questions for each topic. The questions are sequenced to provide smooth transitions throughout the interview (Labaw, 1980). Development of questionnaire items is based on hypotheses that the researcher has developed. The hypotheses are presented to a group of individuals who are subject matter experts, and they perform a preliminary assessment of the hypotheses (Labaw, 1980). The questionnaire may require modification if the hypotheses are not viable.

Most survey formats that gather data through the interview technique would not use response alternatives that are dichotomous, such as a "yes/no" response. There is not much to probe with this type of format (Bradburn & Sudman, 1979). There have been exceptions where interviews used a "yes-no" checklist. Krohn (1984) advised using this dichotomous format for a series of separate scenarios that determined human factors problems unique to the task. The checklist was accompanied by comments. Numerous scenarios could be discussed, and interview time was minimized by this procedure. The advantage in using this approach was that this provided a maximum, or at least a satisfactory coverage of the topics during one interview with each test participant. The constraints of this field test did not allow for follow-up time.

Surveys using interviewers require that questionnaire format contain a logical sequence for the interviewer to follow. To insure consistency across interviewers, it is customary to develop instructions for the interviewer regarding how to use the questionnaire form. It is possible to design a questionnaire with interviewer instructions embedded in the body of the questionnaire. These instructions are usually set off by capital letters that have been enclosed in parentheses (Backstrom & Hurchur-Cesar, 1981).

Telephone interviews benefit if the surveys contain items and response formats which differ from face-to-face interviews. To reduce the potential for phone disconnects, questionnaires must have fewer items than face-to-face interviews. These items should be shorter. This facilitates a higher interaction between the respondent and the interviewer. It tends to reduce the number of telephone disconnects. Telephone interviews preclude the use of visual cues that might be found on response cards in face-to-face interviews (Backstrom & Hurchur-Cesar, 1981).

For face-to-face interviews, interview schedules may include items that are both closed and open-ended. As with telephone interviews, all interviewers require training to maintain content validity and reduce potential interviewer bias.

Survey instruments using interviews have received many types of field assessments. Krohn (1984) used interviews in human factors engineering tasks involving recovery vehicles. Interviews were used by Nemeroff and Wexley (1979) to assess performance feedback characteristics, and Vance, Kuhnert, and Farr (1978) used questionnaires subjected to an interview technique as a way to select employees. Interviewing is the most common technique used for national opinion surveys, and large samples are employed as well (Groves, 1979).

Examples of Interviewing

Face-to-face interviews were used by Krohn (1984) as part of a human factors evaluation. Recovery vehicles, tasks, and equipment were subjected to operational testing. Interviews were used in conjunction with "yes/no" checklists. Each checklist was dedicated to a different aspect of recovery vehicle operation, such as: maintain and repair, recover vehicles including fuel transfer, tow vehicles, and amphibious operations. The format provided a space for comments relating to equipment, and tasks in each category of recovery vehicle operation (these were termed scenarios).

The "yes-no" checklist, in conjunction with the comments, serve as cues to the interviewer to probe in depth any safety hazards and the human factors problem areas that the respondent identifies. Krohn (1984) indicated that administration of the checklist and an interview of critical areas would reduce the total interview time. This approach to interviewing would focus on interviewing only in problem areas. A portion of the checklist that Krohn used in conjunction with interviews follows.

SPECIFIC SCENARIO INTERVIEW OUTLINE

TASK: EXTRACT, REPLACE AND TRANSPORT EQUIPMENT

EQUIPMENT PROBLEMS

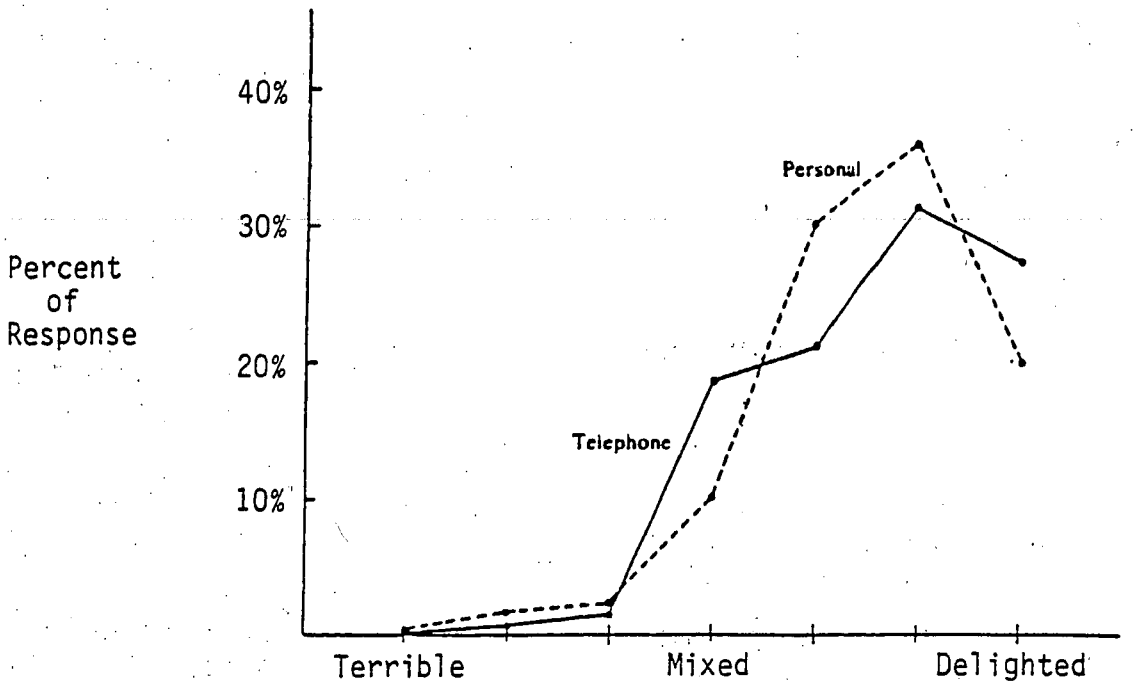
"I will name equipment from the LAVM/RV that you may have used to extract, replace and transport equipment. Please answer Yes or No to indicate whether or not you experienced any difficulties using the equipment. I would also appreciate your comments concerning the difficulties. If you have no experience using the equipment, then state Not Applicable."

Equipment	Yes	No	NA	Comment
1. Crane	___	___	___	_____
2. Crane remote controls	___	___	___	_____
3. Crane onboard controls	___	___	___	_____
4. Winch	___	___	___	_____
5. Winch controls	___	___	___	_____

This checklist interview could serve as the foundation for the generation of another more refined instrument. The checklist interview used by Krohn has the potential to elicit information to use in place of the subject matter expert group (Labaw, 1980). Their functions appear to be somewhat similar.

Research performed by Groves (1979) investigated personal interview surveys and telephone surveys. Response cards used in face-to-face interviews were adapted for interviews conducted over the telephone. A number of response card adaptations were compared. In one research condition, personal survey and telephone survey respondents were questioned about their "life satisfaction" on a "satisfied" to "dissatisfied" scale. Three labeled points were described, and respondents were requested to select a numbered point on the scale. In another condition, questions about "life satisfaction" were presented by the interviewer on a "delighted" to "terrible" scale with seven labeled scale points. Since all scale points were presented by labels instead of by numbers, respondents selected labels for the interviewer to code. Following is a modified version of a chart that Groves used to compare responses by interviewing condition (telephone or face-to-face).

Telephone and Personal Interviews
Delighted-Terrible Scale Response Distribution



Seven-point scale used to measure satisfaction with "Life as a Whole"

Comparisons of Interviewing

Much of the research on questionnaire construction as it relates to interviewing is separated into various subtopic areas. That is, interviewing has been investigated in areas, such as: the impact of ethnicity, interviewer speech behaviors, face-to-face interviews, and telephone interviews.

Difference in response rates is frequently an issue when comparing interview methods. Some investigators prefer interview surveys to other types since they have a history of higher response rates. Orlich (1978) indicated that surveys conducted through face-to-face interviews would produce a 100% response rate. However, the response rate predicted by Orlich may fall well below 100%. Shosteck and Fairweather (1979) found that respondents who were physicians had completion rates of 74% for face-to-face interviews.

In a study conducted by Shosteck and Fairweather (1979), mailed questionnaires were compared to a face-to-face interview. The questionnaires were identical for both conditions. They determined that a mailed questionnaire was superior to face-to-face interviews. It took less time and cost less money ($24 versus $63) to obtain the data. The final completion rate was 70% for the mail survey and 74% for the face-to-face interview.

Weeks and Moore (1981) researched the ethnic background of interviewers. They examined whether the relationship of the ethnic background of the interviewer and respondent would bias the responses on a survey. The ethnic backgrounds of respondents examined were: Cubans (residing in Miami), Chicanos (residing in El Paso), Chinese (residing in San Francisco), and native Americans (residing in Northeast Arizona). There were 101 interviewers used in this study (50 ethnics and 51 nonethnics). Ethnic interviewers were defined as being members of one of the four ethnic groups: Cuban, Chicano, Chinese, and native American. Nonethnic interviewers were Caucasian -- not of Latin descent. The results indicated that there were no significant differences between ethnic and nonethnic interviewers.

For survey items which are non-threatening, and are non-social, the difference in ethnicity between interviewer and respondent does not appear to bias survey results. Weeks and Moore (1981) reported the work of other researchers which supported their results for Mexican American and Anglo interviewers (Welch, Comer, & Steinman, 1973), and caucasian and black interviewers (Hyman, Cobb, Feldman, Hart, & Stember, 1954; Williams, 1964; Schuman & Converse, 1971; Hatchett & Schuman, 1975; Schaeffer, 1980). Ethnicity of interviewer does not appear to bias survey results.

Bradburn and Sudman (1979) strove to eliminate errors in interview-questionnaire administration. They were interested in identifying the interviewer characteristics which contributed to the error. They observed that on about half of all questionnaire administrations, interviewers committed non-programmed behavior and errors instead of faithfully following the interview schedule. They controlled for interviewer errors through stringent selection criteria and interviewer training. They had interviews coded for non-programmed speech behavior on 41,292 individual item administrations. According to their analysis, reading errors occurred more frequently than any other type of error. They were able to identify interviewer and respondent characteristics which contributed to the non-programmed behaviors. Age was the primary variable affecting interviewer and respondent behaviors. These behaviors threatened the standardization for administering the questionnaire. They found that older respondents (65 and over) had a higher interaction rate with the interviewer than younger respondents. Older respondents required prompting to complete the survey items. Older interviewers (55 and over) were less likely to follow the interview schedule as closely as younger interviewers. Therefore, the survey procedures were not as standardized. Non-programmed behavior exhibited by these interviewers may have been due to their greater experience level and less formal approach.

Interviewer characteristics were examined by Groves (1979) for differences between telephone and face-to-face interviews. Groves was concerned with identifying interview strategies (interviewer behavior) that would produce better data collection. He was also concerned with adapting response cards from face-to-face interviews into acceptable communication vehicles for telephone interviews. As with the response rate for face-to-face interviews versus a mail survey (Shosteck & Fairweather, 1979), response rate was higher for the face-to-face interviews than for the telephone interviews (Groves, 1979). Adaptation of response cards used in face-to-face interviews and telephone interviews produced varying results

depending on the type of interview. Labeling the scale points with numbers generated large differences between conditions when the numbers were represented by more uncommon number labeling (Feeling Thermometer number labels ranged from 0 to 100). The response card for face-to-face interviews was not numerically labeled for 75 and 80 degrees. This created the largest single difference between response patterns for the telephone interview and the face-to-face interview. The difference may have been an artifactual one. According to Groves, there were also varying response patterns between interview modes when response cards were numerically labeled in such a way that they were divisible by larger numbers (telephone response cards were labeled 0-50 degrees, 50 degrees, and 50-100 degrees).

The effectiveness of using interviews has received attention for management application. Nemeroff and Wexley (1979) and Kingstrom (1979) investigated this approach for performance feedback. Research findings supported the use of structured interviews for performance feedback. Vance, Kuhnert, and Farr (1978) felt that a structured interview format would be useful for selecting employees. They used behavioral rating scales to compose structured interview items. Psychometric properties of interview ratings were compared for behavioral scales and graphic scales. The hypothesis that behavioral scales would be psychometrically superior was supported. When asked for their preference, managers strongly favored the behavioral scales ($p<.001$). Kingstrom found no significant differences between appraisal format for interviews, and supervisors' willingness to conduct performance feedback interviews. Research results are mixed, and further research is required regarding performance feedback interviews. For employee selection, Vance et al. determined that interview surveys can be reliable. Behavioral scale ratings were significantly more accurate than graphic rating scales.

Research which focuses on interviewing for questionnaires is diverse in content. It is not possible to have implicit confidence in the conclusions about the generalizibility of these findings since the methods, experimental designs, etc. vary so greatly. For example, the subjects used in research on interviews have included persons from non-English speaking backgrounds, managers and subordinates who work in mental hygiene, sales supervisors, members of households across the United States, physicians, and, of course, students.

Conclusions Regarding Interviewing

Research on interviewing techniques that focuses on characteristics of the interview provides evidence that this type of survey is viable when time and money are not a constraint for face-to-face interviews. They are recommended in situations where a high percentage of response rate is valued. Where time is a constraint but cost is not, telephone surveys may serve as an intermediary approach to collecting data. The response rate does not tend to be as high for telephone surveys as that of face-to-face interviews.

Perhaps improving interview techniques that are used for telephone surveys would enhance the response rate. Greater care is required when interview guides are designed for research conditions that include telephone surveys in conjunction with other types of surveys, such as mail or

face-to-face interviews. How the interviewer labels the scale points and how the questions are constructed (unfolding general questions to obtain more specific information) has the potential to bias survey results (Groves, 1979).

The research of Weeks and Moore (1981) supports the contention that a difference in ethnicity between the interviewer and the respondent does not bias the results of the survey. However, in situations where the items are threatening or race-related, the survey results would probably be influenced by ethnicity of the interviewer, as well as that of the respondent. Race-related questions require matching interviewer and respondent by ethnicity. Apparently age is a variable that impinges on the accuracy of surveys which are conducted by face-to-face interviews. Bradburn and Sudman (1979) found that interviewers over 55 years of age tended to present a nonstandardized survey to respondents (they appeared to be less formal in their presentation of survey items). In addition, respondents 65 years and older requested clarification more frequently than younger respondents, and submitted a higher frequency of inappropriate item responses.

When interview surveys have been applied to solve management-type problems, mixed results were obtained. This may be related to the type of interview used. Vance, Kuhnert, and Farr (1978) determined that behavioral scales used in interviews for selection were superior to interviews that incorporated a graphic rating scale. Interviews used for performance feedback conditions were recommended as a way of increasing participation by employees to provide them with an opportunity to set job-specific goals. This type of procedure requires training for supervisors who are performing the interview (Nemeroff & Wexley, 1979). Training interviewers for this task is subject to issues of reliability among format, rater characteristics, and attitude toward feedback interviews (Kingstrom, 1979).

The development of interviewing schedules and training interviewers to conduct standardized interviews is time-consuming. Individuals tasked with conducting military surveys may have underutilized their professional skills in the developmental stages of interview surveys. This is due to the time constraints placed on them (Chun, Fields, & Friedman, 1975). There has been a tendency in military surveys to over sample (large-scale periodic surveys have had response rates ranging between 38% and 51% according to Chun et al.). Military survey research could benefit from the following: exploring various ways to obtain more lead time in survey development, increase response rate, control standardizations in field administration, and control for methodological bias (response bias of respondents brought about by the influence of superiors).

6.2 COGNITIVE COMPLEXITY

Description of Cognitive Complexity

In recent years, researchers have shifted the exploration of scale characteristics, such as number of scale points, to relationships among respondent characteristics, format preference, and other aspects of the rating situation. Originally, the term "cognitive complexity" was defined by Schneier (1977a), although it was developed from Kelly's (1955) theory of personal constructs. Cognitive complexity has been commonly defined as the ability to differentiate person-objects in the social environment.

Cognitive complexity, according to Schneier (1977a), is a trait whereby respondents would have the ability to perceive the behavior of others in a highly differentiated system. It follows then that individuals who are cognitively simple would perceive their environment in a relatively undifferentiated manner (lacking the ability to discriminate between dimensions) (Bernardin, Cardy, & Carlyle, 1982). Out of Schneier's research on cognitive complexity, a theory of cognitive compatibility was formed. Cognitive compatibility purports to enhance the psychometric quality of ratings when the rating format is compatible with the cognitive structure of the respondent (Bernardin, Cardy, & Carlyle, 1982; Lahey & Saal, 1981). Cognitive compatibility theory suggests that cognitively complex respondents should be matched to cognitively complex formats, and that cognitively simple respondents should be matched to cognitively simple formats. It was hypothesized that the matching of respondent to format would increase respondent satisfaction and confidence about their evaluation. The concept of compatibility is especially important since there has been the concern that respondents' ability to discriminate may break down as the number of evaluations they are tasked to make reaches higher and higher levels. The concept of compatibility has been especially important since there is the concern that requesting respondents to make too many evaluations may exceed their ability to discriminate (Jacobs, Kafrey, & Zedeck, 1980).

Examples of Cognitive Complexity

Measures of cognitive complexity were obtained by Lahey and Saal (1981) for participants in their research. Measures were taken on the Role Constructs Repertory (REP) test, factor analysis of the REP test, and a scoring task. These three measures were used to divide scores at the median in order to assign participants to a cognitively complex or cognitively simple designation. Four scale formats were developed, two formats being cognitively complex while two formats were considered cognitively simple. Rating scales used were Behaviorally Anchored Rating Scales (BARS), Mixed Standard Rating Scales (MSS), Graphic Rating Scales (GRS), and an Alternate Scale (AS) with three scale points. They also used a 5-point Likert scale to measure respondents confidence in their ability to make accurate ratings. Following is their description of the four rating scales they used which they considered either cognitively complex or cognitively simple.

"Behaviorally anchored rating scales. The behaviorally anchored rating scales (BARS) contained nine performance dimensions, each of which was rated on a separate 7-point linear scale with both numerical and behavioral anchors. Dimensions contained either 5 or 6 anchors; a total of 50 behavioral anchors appeared on the scales."

"Mixed standard rating scales. After obtaining appropriate behavioral anchors for each level of performance on each of the nine dimensions, three levels were chosen for inclusion in the mixed standard rating scales (MSS); one statement reflected superior performance, one reflected average performance, and the third reflected inferior or poor performance. The statements for the nine dimensions were randomly ordered, and raters were asked to indicate if their instructor was better than, accurately described by, or worse than each of the 17 statements. Numerical ratings for the nine dimensions were determined according to the procedure suggested by Saal (1979), a revision of Blanz and Ghiselli's (1972) original scoring scheme."

"Graphic rating scales. The graphic rating scales (GRS) contained the same nine performance dimensions, and definitions listed on the BARS. The behavioral anchors were replaced with the labels 'exceptionally good' and 'exceptionally poor' at the top and bottom, respectively, of the 7-point numerical scale."

"Alternate rating scales. Adopting the terminology used by Schneier (1977a), an alternate rating scale (AS) was developed by listing the nine performance dimensions, and their definitions, along with a 3-option scale. Raters were asked to place a check mark next to the adjective ("above average," "average," "below average") that best described their instructors' performance on each of the dimensions."

BARS and the GRS were considered the cognitively complex scales, while MSS and AS were viewed as the cognitively simple scales.

Comparisons of Cognitive Complexity

The initial research on cognitive complexity conducted by Schneier (1977a) was initiated out of a concern for the use of BARS. This is due to the fact that when a rating scale has a large number of dimensions to rate, this may impose a cognitive overload on the respondents. In this context, it was felt that they are no longer able to accurately discriminate among dimensions (Jacobs, Kafry, & Zedeck, 1980). Schneier used two formats in his research. BARS served as the cognitively complex format, as well as a simpler format for use by cognitively simple respondents. The subjects in Schneier's research were manufacturing workers (Sauser & Pond, 1981). One outcome of this research was that the cognitively simple raters preferred the cognitively simple form, while the cognitively complex raters preferred BARS. It was found that cognitively complex raters exhibited less restriction of range and less leniency than cognitively simple raters when the BARS format was used. In addition, less halo was exhibited by complex raters regardless of whether the format was complex or simple (Bernardin, Cardy, & Carlyle, 1982).

The intuitive appeal of aligning complexity of format to cognitive complexity of respondent prompted researchers to investigate this phenomenon. The results of further research have been disappointing. Schneier's (1977a) research indicated that the characteristics of the rater may influence the quality of the ratings (Borman, 1979). However, attempts to replicate his findings have not been supported. Bernardin, Cardy, and Carlyle (1982) compared a cognitively complex BARS (MSS were used in one of the experiments) format with a cognitively simple GRS in four different experiments. Their findings indicated that there was no significant relationship between respondent's cognitive complexity and confidence in rating scale, halo, and scale acceptability. None of the four experiments produced any evidence supporting Schneier's theory of cognitive complexity. It has been noted that the conditions of Schneier's original research on this topic has not been exactly replicated. Schneier's subjects were manufacturing workers. Subjects for Bernardin et al. were students in three of their experimental groups, and police sergeants and patrol officers in one experimental group. Schneier had subjects rate 14 dimensions for his cognitively complex format and 10 dimensions for his cognitively simple format. Bernardin et al. varied the number of dimensions measured in the four experiments between 5 and 13.

Sauser and Pond (1981) explored the effects of training and scale construction participation on cognitive complexity. It was hypothesized that psychometric error would be reduced by having raters participate in scale construction or receive training. They used BARS with 9 and 11 dimensions with their student raters. BARS with 9 dimensions was considered simple, and BARS with 11 dimensions was considered complex. Even though the rater groups were significantly different from each other at the .0001 level for cognitive complexity, there were no significant multivariate findings for cognitive complexity x participation x rating (leniency error was not affected by these variables). Their study showed no evidence to support the contention that cognitive complexity, training, and scale construction participation reduced bias and error in ratings.

Using college students as subjects, Lahey and Saal (1981) measured the cognitive demands of rating using a GRS, BARS, MSS, and a 3-point AS. All four scales consisted of nine dimensions. The cognitively complex formats were the GRS and the BARS, each having seven scale points. The cognitively simple scales each had three scale points. They investigated the characteristics of cognitive complexity as they relate to psychometric quality. They found no significant differences, either as a function of cognitive complexity or an interaction for cognitive complexity x scale format (leniency, halo, and range restriction). Cognitive compatibility as a theory was not supported across four different rating scale formats, and across three different operational definitions of cognitive complexity.

The research performed to investigate cognitive complexity indicates that the variables which identify respondent charactistics for cognition as they relate to scale format are not currently known (Bernardin, Cardy, & Carlyle, 1982; Sauser & Pond, 1981; Lahey & Saal, 1981).

Conclusions Regarding Cognitive Complexity

Reviews of performance appraisal literature for future research in scale construction have suggested that cognitive complexity may be an important rater characteristic. Yet, several attempts to replicate Schneier's (1977a) findings have been to no avail. Several suggestions have been made as to why it has not been possible to substantiate the cognitive complexity hypothesis.

Sauser and Pond (1981) mentioned an explanation attributed to Bernardin and Boetcher (1978) where there is the possibility that for cognitive complexity to be a meaningful research variable, it would require scales that are composed of more than seven dimensions of performance. Another discrepancy is comparing research on cognitive complexity with the 1977 work of Schneier. Most subsequent studies were performed with students rather than with workers in manufacturing plants. Instead of rating peers (in manufacturing plants), students rated their professors (Sauser & Pond, 1981). Lahey and Saal (1981) suggested that Schneier's BARS may have been too complex for most practical situations.

For whatever the reasons, researchers have not been able to provide evidence that cognitive complexity is an important variable in rating behavior. The continuous failure to replicate Schneier's findings casts doubt on the validity of cognitive complexity as an issue.

6.3 EDUCATION

Description of Education

Education, as it relates to questionnaire construction, usually means the educational level of the respondents in conjunction with other demographic characteristics, such as: gender, age, and ethnic background. The effect of education on respondent ratings is frequently examined by researchers conducting surveys (Schuman & Presser, 1981; Messmer & Seymour, 1982; Smith, 1981).

There have been several approaches to examining the influence of the educational level of respondents on the way they respond to rating scales and which type of scale they prefer. An illustration of this research is that done on the relationship of educational level to the use of the "don't know" response over a wide range of issues (Schuman & Presser, 1981). Response consistency, over time, has been examined for its relationship to educational level (Schuman & Presser, 1981). Respondents' preference for different types of scales has also been investigated as a function of educational level (Lampert, 1979). There is the possibility that respondents' rating of items may, in other respects, be a function of their educational level (Smith, 1981).

Examples of Education

To collect information on the educational level of respondents, researchers have had to define educational categories for purposes of collecting data.

In General Social Surveys for 1976 and 1978, respondents were found to frequently ignore items with absolute phrasing. Further responses on subsequent items became contradictory (Smith, 1981). Because of this trend, several hypotheses were formed. Smith hypothesized that subjects who rated questionnaire items with contradictory response patterns would have lower education and/or lower intelligence. It was suggested that these subjects would misunderstand the questions due to cognitive limitations. These limitations would be manifested as: lack of imagination to visualize a range of situations, and lack of experience with questionnaires. This hypothesis was examined by determining the years of schooling that the respondents had completed. In addition, a 10-item word identification test was used to measure respondents' verbal ability. Interviewers also evaluated the comprehension of the respondents.

Response rates in survey research were investigated by O'Neil (1979). Using random digit dialing, a general population telephone survey was conducted for over 1,200 households in Chicago. The effects of respondents' refusals were investigated by placing up to 20 call backs at staggered times to reach persons who had not previously been at home. Elaborate follow-up strategies were used to persuade respondents who initially refused to participate in the survey. Individuals conducting the survey tried to persuade them to change their minds and participate in the survey. O'Neil used many demographic characteristics for this study, including

educational level. Listed below are the demographic categories used to describe this survey sample:

Occupation	Family Income
Age	Religion
Ethnicity	Children
Race	Owner Occupancy
Education	Dwelling Type

The interpretation of the results of this survey has been difficult because the selection of respondents within household was not done randomly. The random digit dialing procedure produced a sample that was divided in half by randomly selecting half of the respondents in the sample to attempt an interview with a male respondent. The other half of the sample consisted of whoever answered the phone first. This resulted in nonrandomization within household selection for the demographic characteristics of education, age, and occupation. The other demographic characteristics listed above were considered to almost never vary within households.

It is not unusual to ask respondents more than one question regarding their educational level. For example, Survey Research Center (SRC) asks respondents to identify, by year, their highest grade of school or year of college completed up through grade level 17 and beyond. They then request additional information from respondents. Respondents are questioned about whether they obtained a high school diploma or passed a high school equivalency test in lieu of a diploma. Respondents are also queried about whether they have a college degree (Schuman & Presser, 1981). Education has been an important variable in describing the demographic characteristics of a sample. However, it is always used in conjunction with other variables.

Comparisons of Education

Various assumptions have been made about the educational level of respondents, and the relationship of this characteristics for its effect on rating items. Studies dealing with the educational level of respondents were designed to answer research questions about other questionnaire construction topics, such as branching, scale preference, and "don't know" response. Educational level of respondents is commonly measured, along with other demographic characteristics, for gender, age, and ethnic background. This is to determine the effects on responses of this variable by itself, or in combination with other demographic variables.

Presser and Schuman (1980) compared the effects of omitting or offering a middle alternative in forced-choice attitude questions on five experiments. In addition to the replication on several national surveys, one of their interests was whether education was related to rating the middle alternative. They were also interested in effects of educational level on omission of the middle alternative. They hypothesized that education would be related to these form differences. They felt that respondents with less education would be the most influenced by inclusion or omission of the middle alternative. They were not able to support this hypothesis. Evidently, education is not related to how respondents use the middle alternative. Responses by educational level do not appear to be affected by whether or

not the middle alternative is included or omitted. Offering the middle category, of course, increases the number of responses to that category, but does not appear to affect meaningfully the overall distribution of responses.

It was determined by Schuman and Presser (1981) that the "don't know" response alternative was selected most frequently by respondents who had the least amount of education. Individuals with less education appear to be those respondents who are most influenced by format when a "don't know" response is included. However, for survey items which request an opinion for an obscure topic area, there is a propensity for respondents with higher levels of education to select a "don't know" response. Respondents with less education have a tendency to give an opinion. Respondents with higher levels of education seem to be more willing to admit they do not have knowledge of a topic area. Respondents with low levels of education do not appear to admit they don't know. Instead, they select a response alternative to represent their opinion.

In a General Social Survey, Smith (1981) found evidence that respondents had a tendency to ignore the absolute phrasing of the survey items. This caused contradictory response patterns. In one of Smith's hypotheses, it was suggested that respondents who had lower education/ intelligence would misunderstand the general questions. This would produce contradictory response patterns. Smith found that respondents with contradictory response patterns had significantly less education, and lower comprehension associated with lower verbal achievement. These respondents were more likely to be non-caucasian and female. The items used in the questionnaire were related to the approval of hitting by private citizens and police. Respondents with contradictory response patterns were less in favor of punitive actions than other respondents. Perhaps researchers need to evaluate general-type questions which appear to result in ambiguous meaning. This would be especially important for respondents with lower educational levels. The items with absolute phrasing were not answered as though they were absolute, but instead they were answered as though they were nonabsolutes. For example, Smith used an absolute question as follows: "Are there any situations you can imagine in which you would approve of a policeman striking an adult male citizen?" Even though some respondents answered "no" to this question, they subsequently approved of situations where they accepted the use of physical force by a police officer or citizen. Respondents with higher levels of education appeared to be able to understand the phrasing and meaning of the questions which used absolute terms. Therefore, their responses were not contradictory.

Demographic characteristics were measured by Messmer and Seymour (1982) in their research on item nonresponse. They examined responses of a large sample (2,114 respondents) in a mail survey for items which immediately followed a branch. They hypothesized that: "The frequency of item nonresponse will be greater for questions immediately following a branch instruction than for those questions which do not follow a branch." Out of the eight hypotheses established by Messmer and Seymour, two were directed toward the demographic characteristics of education and age. The hypotheses relating to education and age are presented here: "The greater the level of education of the respondents, the lesser the frequency of item nonresponse for branching items." "The greater the age of the respondent, the greater the frequency of item nonresponse for branching questions."

The Messmer and Seymour (1982) hypothesis for the adverse influence of item nonresponse for questions immediately following a branch was supported at the .005 level of significance. These hypotheses were partially supported. Hypotheses regarding the demographic characteristics for age was supported, but not for education. They did not reach a level of significance for education, as it is associated with the frequency of nonresponse for items immediately following a branch. However, item nonresponse increased for older respondents (e.g., 60 and older). This was significant at the .014 level. Item nonresponse did not seem to be influenced by education, gender, distance from the branching question to the resulting question, number of previous branches, branches which deal with future behavior, or branches which require an attitudinal response.

The work of O'Neil (1979) for nonresponse to telephone surveys is expanded upon here. Obviously, it is a difficult task to describe subjects who refuse to participate in a survey. Yet, O'Neil tried to identify the characteristics of these subjects. The effects of nonresponse under varying conditions was studied. It was determined that respondents with less education and lower incomes tended to initially refuse to participate in the survey. Other demographic characteristics described these individuals as more likely to be over 65 years old, caucasian, and of Polish, German, or Irish descent. Researchers must decide how much extra expense they are willing to incur to minimize nonresponse rate, and what benefits they derive by increasing response rate. O'Neil was able to increase the response rate up to 86.8% from the initial response rate of 74.5%. The original sample, prior to telephoning respondents, consisted of 1,392 eligible households.

Education was used as a demographic variable along with gender, age, and ethnic origin to investigate the respondents' ability to indicate their attitudes on four different types of scales (Lampert, 1979). Education was the only variable which differentiated among the subjects. Respondents with education below the grammar school level were significantly different in rating the scales at the .0001 level of significance than the other respondents in the sample. Respondents with educational levels above partial high school education did not influence the correlation coefficients. This was because their response distribution was not much different from the distribution of those subjects whose educational level was even higher. The four scales employed were: Attitude Pollimeter (an attitude continuous scale with visual elements), as well as a verbal, numerical, and continuous bipolar scale (see Section 2.6, Continuous and Circular Scales).

This research indicates that, for some studies, educational level of the respondents may be associated with how items are rated. Research findings do not consistently support this contention, although there is partial evidence.

Conclusions Regarding Education

Some evidence supporting the hypothesis that education may influence response patterns was presented by Schuman and Presser (1981) regarding the omission or the use of the "don't know" response alternative. Evidently, respondents with a low level of education were most influenced by format since they had a tendency to select the "don't know" response alternative more frequently than respondents with higher levels of education.

For the purposes of their research, education was differentiated into three levels. They collected very specific data as to the number of years of education completed by respondents. A high level of education was defined as "some college." A middle level of education was considered to be those individuals who had never been to college but had completed a high school diploma or equivalency certificate. Respondents who did not have a high school diploma or who did not pass a high school equivalency test were identified as having a low educational level.

For survey items about obscure topics, there was a greater tendency for respondents with a high level of education to rate the items with a "don't know" response. Those respondents with a low educational level tended to select a response alternative that represented their intensity of emotion in cases where they probably had no opinion (for items they apparently knew nothing about). This research included the general population of the U.S. as the sample. It would be useful to know whether these same characteristics can be generalized to enlisted personnel who have low levels of education. If this were true, then it might be useful to omit the "don't know" response alternative for topic areas that the respondents have experienced. Items should be reviewed to ensure that the respondents understand the content so that they don't select an opinion on a scale just to appear knowledgeable.

Schuman and Presser (1981) found an interaction among educational level of respondents, their response consistency in rating items, and their intensity of feeling about the item. There appeared to be a greater consistency in rating items when respondents had a middle and higher level of education, and had intense feelings about the content of the item. This interaction was not significant for individuals with a lower level of education. They suggested that respondents with a lower level of education may have a more difficult time separating out their attitude toward the content of an item and their attitude strength as a personal response style.

In related research conducted by Smith (1981), respondents with a low level of education misunderstood items which had been phrased using absolute terms. Because of their faulty interpretation of the items, their response patterns were not consistent. The response patterns were contradictory. Smith could have reworded the general questions to avoid confusion by respondents who have low levels of education. Research by Schuman and Presser (1981) and Smith (1981) indicates that respondents with lower levels of education may not understand the content of an item in the same way as other respondents with higher levels of education. Respondents with varying levels of education may not be interpreting the items in the same way.

There is some evidence that respondents may be marking items and the "don't know" response alternative in divergent ways (based on their educational level). The actual format of the survey has the potential to contribute to divergent ratings, too. Using education, gender, age, and ethnic origin as demographic variables, Lampert (1979) found that the educational level of respondents was influenced by format. Results indicated that respondents with a low level of education (in this instance low

level refers to education below the grammer school level) were significantly different than the sample as a whole in rating four different scales. Above the grammar school level, there were no significant differences among groups for any of the demographic characteristics.

In some instances, education appears to be linked to nonresponse. In a telephone survey by O'Neil (1979), low educational level was associated with respondents who had initially refused to participate in the survey (these individuals would have probably been nonrespondents had elaborate follow-up procedures not been used with them). In this particular study, education was related to other characteristics of age, race, and ethnic origin. Education is, at times, the only demographic characteristic obtaining levels of significance, although education is usually found to be associated with other variables.

There have been occasions where researchers felt that education may have influenced the outcome of questionnaire responses, yet this phenomenon was not consistently supported by psychometric results. For example, Messmer and Seymour (1982) hypothesized that education would be a factor in item nonresponse for items immediately following a branch. Results did not reach a level of significance to support this hypothesis. They did determine that age was associated with item nonresponse. Presser and Schuman (1980) compared the omission or the inclusion of a middle alternative in forced-choice attitude questions. They felt that education would be associated with these form differences. They were not able to obtain evidence that would support their contention. Very little is known about the effect of educational level on responses to questionnaires that are designed for use in performance appraisal. In a review of the literature on performance appraisal, Landy and Farr (1980) reported the work of Cascio and Valenzi (1977). They investigated rater educational levels on supervisory ratings for job performance of police officers. Education contributed only a small percentage of the total rating variance, and this finding was not considered to be of practical importance. The demographic characteristic "education," at times, impacts on questionnaire construction. Education is usually linked to other demographic characteristics when it affects response patterns.

6.4 ETHNIC BACKGROUND

Description of Ethnic Background

Survey research investigations into ethnic background have usually been focused on differences between black and caucasian respondents (Landy & Farr, 1980; O'Neil, 1979). There have been exceptions to this trend where research in ethnicity has been expanded to include Cubans, Chicanos (and other Hispanics), native Americans, Chinese (and other Orientals), as well as those of Polish, German, and Irish descent (Weeks & Moore, 1981; Imada & London, 1979; O'Neil, 1979).

Ethnicity has been examined from a number of perspectives, such as: ethnic background of the interviewer, ethnic background of respondents, culture-free content of questionnaire items, effects of ethnicity on performance appraisal scores, response rates associated with ethnicity, use of self-assessment instruments, and surveys of race relations (Segal & Savell, 1975; van Rijn, 1980; O'Neil, 1979; Landy & Farr, 1980; Schuman & Presser, 1981).

Examples of Ethnic Background

Research on ethnic influences on surveys has been quite diverse, as will be apparent from the examples below. Investigations have focused on such issues as differences between black and caucasian interviewers, the ethnic background of individuals who refused to be respondents in a telephone survey, and the extent to which stereotypical ratings are a function of the rater, the scale, and the stimuli being rated.

Schuman and Presser (1981) conducted an experiment to measure response differences obtained by black or caucasian interviewers. They asked respondents the following question:

> "Tell me who two or three of your favorite actors or entertainers are?"

The responses were later coded according to ethnic background of respondents and interviewers to determine the differences.

O'Neil (1979) contacted 1,209 Chicago households to identify individuals from different ethnic backgrounds who refused to be part of a telephone survey prior to extensive follow-up techniques. Questions were asked relating to neighborhood crime. O'Neil eliminated ethnic groups who comprised less than 4% of the sample. To determine the ethnic background of respondents, they were specifically asked:

> "What foreign country would you say that most of your ancestors come from?"

Semantic differential scales were developed by Imada and London (1979) to measure ethnic stereotypes. Respondents received a questionnaire consisting of a biographical information form and a page of instructions. A

3-way interaction among the scales, stimuli, and subjects was analyzed. Subjects were caucasians, blacks, and Orientals. Their social perceptions of ethnic stereotypes were measured on an 8-point bipolar scale with 24 sets of adjectives. The adjectives they used are listed below:

headstrong-mild, gentle
excitable-calm
simple, direct-imaginative
careless-fussy
pessimistic-optimistic
undependable-responsible
uncooperative-cooperative
aimless-motivated
irritable-good natured
maladjusted-adjusted
unsuccessful-successful
quitting, fickle-persevering
disreputable-reputable
nervous-poised
clumsy-refined
silent-takative
shy-outgoing
secretive-frank, open
static-dynamic
submissive-dominant
passive-active
weak-strong
insensitive-sensitive
powerless-powerful

Research in ethnic background effects has been quite diverse for various aspects of ethnicity, including: scale design, ratings, content, and implementation. Studies have focused on different ethnic backgrounds of interviewers and respondents.

Comparisons of Ethnic Background

Implications exist for biasing survey results due to ethnicity whenever surveys incorporate face-to-face interviewing. This potential for biasing could be due to the different ethnic backgrounds of interviewers and respondents. Also, the content of the survey items may possibly produce biased results.

Schuman and Presser (1981) addressed these issues by conducting a study where the race of the interviewers (black and caucasian) was varied along with the context of the questions. One of the items on the survey asked respondents to name their favorite entertainers. The researchers used two survey forms. Each form included questions related to racial discrimination. They hypothesized that more black entertainers would be identified on the first form than on the second form. The first form started out by asking racially-discriminating questions, and then asked respondents to identify a favorite entertainer. The second form asked respondents to identify a favorite entertainer. This request was followed by racially-discriminating questions. The respondents were black, and it

was hypothesized that more black entertainers would be identified when interviewers were black than when interviewers were caucasian. There was no significant difference for question order effect. Racially discriminating items before or after a request to name favorite entertainers apparently did not influence the responses. There was no significant effect for race of interviewer.

Investigation of the ethnicity of interviewers was conducted by Weeks and Moore (1981) in a sample of 1,472 household respondents from non-English language backgrounds (Cubans, Chicanos, native Americans, and Chinese). They analyzed whether there were any significant differences in interview results by the ethnic interviewers mentioned above and Anglo-American interviewers. Surveys and test scores were compiled for each subject. They found that there was no significant difference in interview results between ethnic interviewers and Anglo-American interviewers. This research supports previous findings for interviewer effects for black and caucasian interviewers. Interviewer nonprogrammed behaviors that include reading errors, feedback, and requests for clarification were investigated by Bradburn and Sudman (1979) for possible ethnic influence. They determined that there were no significant differences for race or sex of interviewer, although age of interviewer was a factor.

In a telephone survey of 1,209 Chicago households, O'Neil (1979) sought to isolate the characteristics of individuals who refused to participate in a telephone survey. Individuals initially not willing to participate in the survey tended to have a lower income, were older, and had less education than other members of the sample as a whole. For this survey, resistant respondents were caucasian blue collar and service workers of Polish, German, and Irish descent. (Black respondents, overall, did not have a tendency to refuse participation in the survey.)

The ethnic background of raters assessing ratees in performance appraisal scales was reviewed by Landy and Farr (1980). Results for this type of research have been mixed so that there appear to be no clear guidelines. For example, they reported the work of Crooks (1972), De Jung and Kaplan (1962), and Hamner, Kim, Baird, and Bigoness (1974) where higher ratings were received by ratees when the rater was of the same race. They also reported the work of Schmidt and Johnson (1973) and Bass and Turner (1973) where there was no significant effect for same-race raters with peer ratings, and no significant differences for black and caucasian raters.

Self-assessment has been used by ratees in essentially two types of applications. Self-assessment has been used by applicant candidates for selection into new positions. In addition, self-assessment has been used by individuals in performance appraisal. Ratees appraise their own performance instead of being evaluated by a supervisor (van Rijn, 1980). Van Rijn reviewed research performed by Levine, Flory, and Ash (1977) where self-assessment for minority group members for typing abilities were examined. Self-assessment ratings were somewhat similar for caucasians and minorities. However, caucasian job applicants were able to predict their typing scores at $r = .64$, while minority applicants predicted their typing scores at $r = .39$. In the range of self-assessment ratings, the common regression line for the total group underpredicted the performance of

caucasian applicants, and overpredicted the performance of minority applicants. In a study reported by van Rijn (1980) and conducted by Hardt, Eyde, Primoff, and Tordy (1978), van Rijn indicated that applicants rating their knowledge, skills, and abilities for police officer positions had a low correlation with actual tests. There were differences among caucasian, Hispanic, and black applicants; blacks and Hispanics were apparently not as aware of their abilities, knowledge and skill level as were caucasiana applicants.

The impacts of ethnicity on the reliability and validity of self-assessment ratings requires further research. Research findings indicate differences in responses based on ethnic background. Bacause of the paucity of research in this area, it is premature to infer trends for ethnicity in relationship to self assessment.

Response patterns for caucasian and non-caucasian high school students was investigated by Arima (1980). The Armed Services Vocational Aptitude Battery (ASVAB) was compared with a performance-based, culture-free test. There were no significant differences between caucasian and non-caucasian male subjects, as well as no significant differences between caucasian male and caucasian female subjects. There was a significant difference between caucasian and non-caucasian female subjects. According to Arima, female respondents typically scored lower on the ASVAB. Evidently, females show equal scoring to males on the clerical tasks only. Research performed by Schmidt and Hunter (1980) indicated that tests which appeared to be more valid for one race than for another were not really able to meet the psychometric rigor of validity. Sedlacek (1977) pointed out that evidence does not support content making a difference for scoring on tests for racial minorities or females. Administration may be what is influencing the results. It is known that the ethnic background of a test administrator, and the perception by the respondent for the use of the scores, has the potential to influence test results. Sedlacek was referring to the language used in the items. In research performed by Arima, item content did influence scores on vocational tests. For females (caucasian and non-caucasian), there were scoring differences. This may have been due to the vocational content of the tests.

The U.S. Army has conducted surveys on race relations (Segal & Savell, 1975). Collecting multiple sources of data that accurately reflected ethnicity within the Army was approached by supplementing surveys with field observations, intensive interviews, Army records, and experimental programs. Even though alternate methods for collecting data were used, surveys were considered the primary data collection instrument. They suggested that a better understanding of questionnaire construction methods and sampling theory would improve the quality of data.

The effect of ethnic background for the areas of self-assessment, performance appraisals, and vocational tests are mixed. Studies provide little guidance for the structuring of item content on questionnaires in these areas. Research associated with interviewer effects may be more useful in application.

Conclusions Regarding Ethnic Background

There has been a consistent trend related to ethnic backgrounds of interviewers and respondents. Studies indicate that nonsensitive, non-racial items appear to be relatively immune to interviewer effects for ethnic background (Weeks & Moore, 1981; Welch, Comer, & Steinman, 1973; Hyman, Cobb, Feldman, Hart, & Stember, 1954; Schuman & Converse, 1971; Schaeffer, 1980). Schuman and Presser (1981) found no significant effect for ethnic interaction with the interviewer and respondent. This research was supported by the work of Weeks and Moore (1981) and Bradburn and Sudman (1979), where ethnic background of interviewers was compared. There was no significant difference in interviewing by race. Response rates for surveys were examined for ethnicity by O'Neil (1979). Even though ethnic backgrounds were identified as Polish, German, and Irish for those termed resisters (subjects which did not want to participate in the survey), the real issue may be one of age and socio-economic background. For example, in searching for characteristics which explain interviewer errors, Bradburn and Sudman (1979) discovered that age was the only factor identifiable while ethnic background was not an issue.

Scales developed for performance appraisal, when implemented, have at times been subject to the effects of rater bias due to the ethnic backgrounds of raters and ratees (Crooks, 1972; De Jung & Kaplan, 1962; Hammer, Kim, Baird, & Bigoness, 1974; Landy & Farr, 1980). However, there have been inconsistencies in the research so that some researchers have not been able to replicate the effects of ethnic background on rating (Schmidt & Johnson, 1973; Bass & Turner, 1973; Landy & Farr, 1980). Performance appraisal scales which were self-administered for selection purposes indicated that caucasian subjects underpredicted their performance. Minority subjects over-predicted their performance (Levine, Flory, & Ash, 1977). A similar finding was observed for self-administered performance appraisals where caucasian subjects were more accurate in evaluating their performance than minority subjects (Hardt, Eyde, Primoff, & Tordy, 1978). Most of the studies on self-assessment have been for lower level jobs, and the results have been mixed. Further research will be required to determine the extent to which self-assessment can be used. Historically, this approach has been fraught with technical and practical problems. It has been suggested by van Rijn (1980) that self-assessment in personnel selection be used to supplement more traditional instruments. Researchers who are involved in questionnaires designed for more traditional approaches to performance appraisal may find it useful to control for possible bias in ratings caused by ethnic backgrounds of raters and ratees.

Questionnaires designed for tests have been subject to ethnic differences (Arima, 1980). Controlling for these differences by changing the wording of items does not necessarily result in modifying response patterns (Sedlacek, 1977). Employment tests which purport to be valid for caucasian and not for black respondents, or valid for black and not for caucasian respondents, have not been psychometrically supported according to Schmidt and Hunter (1980). This places the entire concept of culture-free tests in question. For example, investigating culture-free, performance-based tests with the ASVAB indicated that there were no significant differences between caucasian and non-caucasian samples overall. However, females were negatively affected by these instruments (Arima, 1980). Females performed

poorly on the trade tests. Their performance was equal to that of males in the areas of attention to detail and numerical operations. These areas were both elements of the clerical test. This adversely affected their selection into the technical courses. It is known that there are ethnic differences relating to culture-free tests. Yet, these tests may not be any more valid than tests which were designed on the basis of culture.

This conclusion represents the on-going, broad-based research that is required in questionnaire construction to resolve technical problems associated with ethnic background. The U.S. Army used a sound approach in designing data collection techniques on the topic of race relations. To supplement survey data, a composite of methodologies was used. This provided a method for supplementing and cross-checking the survey data (Segal & Savell, 1975). Not enough is known regarding the impact of ethnic background on questionnaire design. It is possible to reach some conclusions regarding interviewer interactions.

6.5 GENDER

Description of Gender

When questionnaire items are constructed, investigators tend to assume that item content and item order are not affected by respondent gender. Yet, it is not always possible to make such assumptions. Schuman and Presser (1981) analyzed differences in response pattern associated with gender. They compared male-female responses to open-ended items and closed-end items on two different questionnaires. The topic area covered on the questionnaires dealt with job preference. Males and females had different response patterns, although the statistical findings were not highly significant. Females tended to select pleasantness, and males tended to select autonomy as preferred job attributes.

In the study of questionnaire construction, gender of the respondent does, at times, influence the survey outcome. Item content has a potential for item-gender interaction. The ordering of items may also interact with gender. It has been suggested by McFarland (1981) that question order be carefully planned. McFarland examined question order effects for general and specific survey items to investigate the strength of order effects associated with gender and/or education.

Questionnaire construction research has been sensitive to the potential for stereotypical response patterns brought about by rater/ratee item-gender interaction on performance appraisals. Different male versus female response patterns have been found in testing. This is especially true for items in the vocational domain. Whenever a questionnaire is constructed, there is the possibility that item content or item order may produce response pattern differences attributable to gender.

Examples of Gender

The search for response patterns reflecting gender differences has been pursued through the use of various research designs, populations, scaling variations, etc. Many hypotheses have been constructed to explain the response patterns of females and the response patterns of males. The semantic differential is one of the instruments which has been subjected to this type of research.

The underlying three dimensions of the semantic differential (according to Osgood, Suci, & Tannenbaum, 1957) are: 1) evaluative, 2) activity, and 3) potency (see Section 2.3, Semantic Differential Scales). Benel and Benel (1976) reported the hypothesis of Meisels and Ford (1969) and Miller (1974) that differences which occur within the evaluative dimension are attributable to females. It was felt that females have a greater need for social approval than males. According to one hypothesis termed the "impulsivity hypothesis," females will have extreme ratings on all factors in relationships to the male mean's midpoint scale score. Benel and Benel selected emotionally charged concepts to accentuate the impulsivity of

female responses. Listed below are the concepts they identified for their semantic differential.

"Love, Life, Truth, Vomit, Pollution, Beggar."

In a study on question order effects, McFarland (1981) suggested that the order of items on a survey may be critical, depending on the population being surveyed. McFarland sought interactions between question order, education, and gender for a telephone survey of Kentucky households. The survey items were administered using two different forms. On the first form, a series of specific questions was followed by general questions. On the second form, general questions were followed by a series of specific questions. An illustration of general questions used by McFarland is included here.

1. "How would you describe the current energy problem in the United States?"

   a. Extremely serious
   b. Somewhat serious
   c. Not serious at all

2. "During the next year, do you think the economy..."

   a. Will get better
   b. Will get worse
   c. Stay the same

3. "In general, how interested would you say you are in politics:"

   a. Very interested
   b. Somewhat interested
   c. Not very interested

4. "In general, how interested would you say you are in religion:"

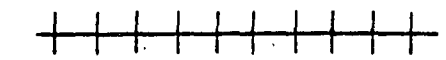
   a. Very interested
   b. Somewhat interested
   c. Not very interested

Research performed by Norton, Gustafson, and Foster (1977) focused on rater-ratee sex bias in rating scales used to measure managerial performance. They indicated that scales should be general enough to describe behavior in varying management situations. These scales need to remain unidimensional to describe only one kind of behavior (construct). Following is an example of a scale item used to measure management skills in setting and achieving objectives. Managers rated male and female versions

on two case histories. They were each identical in content with the exception of minor differences in career histories.

"Has difficulty defining objectives. Sets objective levels which may be unrealistic. Has difficulty establishing priority of needs and may distribute resources inefficiently. Requires extensive supervision to accomplish objectives."

"Defines realistic objectives. Sets realistic objective levels. Ranks the objectives and distributes resources according to the needs of the company. Accomplishes objectives without the need for excessive supervision."



LOW HIGH

Comparisons of Gender

In investigating gender differences for respondent rating patterns, researchers have approached the topic area in divergent ways. Some researchers have measured variations in responses for question form effects (Schuman & Presser, 1981; Smith, 1981; Benel & Benel, 1976) (e.g., comparing open-ended items and closed-end items. Measuring question order effects has been another approach (McFarland, 1981). The difference between male and female response patterns has been especially relevant to those individuals responsible for constructing questionnaires which are used for performance appraisal (Landy & Farr, 1980; Rose, 1978; Norton, Gustafson, & Foster, 1977; London & Poplowski, 1976). Those researchers developing test items for vocational tests have been sensitive to this issue (Arima, 1980).

Regardless of the content of a survey, education and gender are two of the most frequently measured variables. In the study of how items are constructed, there is the assumption that respondents with more education will be less effected by the structuring of the item. In addition, the rating of items by gender may be dependent on the content of the item. This may be related to the value structure that is embedded in the content of the item. Schuman and Presser (1981) designed and implemented a study on work values using open-ended and closed-end items on two different forms. They found that response pattern by gender produced similar patterns for both open-ended and closed-end forms of questions. Females were more likely to select the work value "pleasantness," while males were more likely to select the work value "autonomy." Males selected "security" and "pay" more frequently on the open-ended form than on the closed-end form. However, this finding was only significant at the .10 level. The trend to select "security" and "pay" on the open-ended form was not replicated in further studies.

In an experiment conducted by Smith (1981), response patterns were related to general questions and to specific situational questions. Gender and ethnic background were investigated for their influences in response patterns. Respondents had difficulty with the general questions which were ambiguous and abstract. The general questions resulted in item ratings

that had contradictions in response patterns. Respondents who rated items in a contradictory pattern tended to be female, non-white with less education. McFarland (1981) was interested in question order effects for both gender and education. It was hypothesized that general and specific questions on topic areas would be either diminished or enhanced by placing content specific questions either before or after the general questions. McFarland found that there was no evidence for question order effect by gender or education. It was suggested that specific questions are less prone to order effects than general questions. Perhaps McFarland's general questions were not as abstract as the general questions developed by Smith.

A semantic differential scale was used by Benel and Benel (1976) to investigate male/female differences in rating. It was hypothesized that females would rate items according to social desirability. Their results indicated that male-female ratings were consistent. There were no significant differences between ratings by gender. For the evaluation dimension of the semantic differential (see Section 2.3, Semantic Differential Scales), differences in gender rating have been attributed to the females' need for social approval. For the dimensions of activity and potency, differences in gender ratings have been attributed to a greater impulsivity of females. Neither of these hypotheses was supported.

The effect of gender differences in response pattern have little support. Rating characteristics identified by gender alone are not enough to explain rating differences. Other variables must be taken into account as well. When other characteristics are included with gender, such as ethnic background and education, there is a greater potential for interactions which influence rating. Schuman and Presser (1981), McFarland (1981), and Benel and Benel (1976) were not able to attain a level of significance which would attest to response patterns by gender. Smith (1981) did obtain evidence of a response pattern by gender, but it was combined with other variables that included ethnic background, lower education, lower comprehension, and lower verbal achievement.

Questionnaire construction for items used in performance appraisal has been suspect for differences in response patterns by gender. The perception exists that items may be rated according to sex-role stereotypes. Landy and Farr (1980) reported the work of Schein (1973) where male and female raters held common sex-role stereotypes. For example, male and female managers perceived successful middle-level managers as having common traits ascribed to males. Rose (1978) examined ratings by gender according to attribution theory. The results of this research indicated that male and female raters attributed greater effort and higher ratings to those managers whose subordinates were of the opposite sex. Findings indicated that when subordinates and managers were of the same sex, their managerial performance was rated lower. The subjects used in this research were upper division and graduate students enrolled in business courses. Landy and Farr indicated that there needs to be more performance appraisal research conducted in the actual work environment. Results obtained by Rose were unusual. When gender differences in ratings do occur, they usually take place along the lines of sex-role stereotypes. Male managers tend to receive higher ratings regardless of the gender of the subordinate (Schein, 1973).

Norton, Gustafson, and Foster (1977) trained managers using a case study method. These managers worked for a public utility company. They concluded that there were significant differences at the .01 level for means and variances of ratings for male and female subjects. They found no interaction effect before or after training, or between sex of rater and sex of ratee. London and Poplowski (1976) obtained conflicting results to those of Norton et al. Ratings by females were significantly more positive than ratings by males. The subjects in the London and Poplowski study were students. There may be a difference in performance appraisal ratings by students compared with individuals who are actually on the job. Of course, there are student differences for ratings obtained on vocational-type tests, too (Arima, 1980).

Conclusions Regarding Gender

Review of the research associated with questionnaire construction and difference in response patterns by gender has received mixed results. Some studies have found differences in rating by females and males, while other studies have not. When researchers analyze their data for interactions with other demographic characteristics, there is a greater possibility of identifying gender as an interacting variable. Common characteristics found to interact have been gender, education, age, and ethnic background (Schuman & Presser, 1981; Smith, 1981; Landy & Farr, 1980).

The actual content of an item may elicit rating differences by males and females. For example, items about the work environment in an opinion survey were found to be rated differently at the .10 level of significance (Schuman & Presser, 1981). Arima (1980) found that females taking military vocational tests performed more poorly than males. When standardized norms were applied to the females, there was not equity in the selection of females into the more desirable technical courses. Even so, females were found to be comparable in general cognitive ability. These gender differences in ratings indicate that the content of an item may have the potential to bias it. There are differences in the values that males and females hold. The content of some items is not equally relevant. The respondent may lack the background and experience required to adequately respond to the item. This is illustrated by the differences in vocational scores where males and females are cognitively comparable. Yet, females usually do not have the background experience to adequately respond to the items. This situation is analogous to the argument over ethnic background and culture-free tests.

To reduce the amount of bias in the content of items, it appears to be beneficial to use more content specific items (see Section 4.2, Wording of Items and Tone of Wording). Item content that is general, and possibly ambigous, has been known to produce survey results which are highly questionable. Item content may effect those respondents with less education more than other respondents (Smith, 1981). The question order should probably proceed by constructing questionnaires with general questions first, followed by more specific questions. Concrete items are less prone to question order effects (McFarland, 1981) (see Section 4.4, Order of Items). McFarland found that the strength of order effects did not vary for gender or age.

Questionnaires developed for use as performance appraisal instruments have been examined for male-female rating differences. There has been concern as to whether raters were rating ratees by stereotypes or by actual behavior. There has been evidence to support both sides of this issue (Landy & Farr, 1980; Schein, 1973; Rose, 1978; Norton, Gustafson, & Foster, 1977; London & Poplowski, 1976). Many studies performed in this area have used college students as subjects. Studies investigating rating by gender for work environments might be best performed in "real world" work situations instead of in classrooms.

Questionnaires which measure differences in rating by gender have been found to use almost every possible format known to researchers (Brannon, 1981). The different formats used to measure gender differences have not all proven to be equally desirable (see Section 7.1, Questionnaire Layout). The issues of response style for males and females, as it relates to format, are no different than the issues of selection of format for other kinds of measurements. The question is not whether males or females will be using the form, but what is the purpose of the study. For example, open-ended questions may be good for an exploratory study regardless of whether the respondents are male or female. The development of sound items following appropriate scale development procedures is the best defense against items which are susceptible to rating differences by gender. If the investigator suspects differences in rating by males and females, then the interaction of other characteristics should also be examined.

6.6 AGE

Description of Age

Demographic characteristics for age and questionnaire construction are usually related to education, and sometimes to ethnic background and gender. These characteristics, in combination or individually, may influence the way in which a respondent rates a scale. Many experiments have been conducted on questionnaire surveys which take these variables into consideration (O'Neil, 1979; Landy & Farr, 1980; Messmer & Seymour, 1982). For example, Messmer and Seymour examined the effect of branching on item nonresponse. The researchers examined each branch to assess whether the respondent correctly followed instructions. The influence of demographic characteristics was measured for age and education of respondents. O'Neil investigated whether response rates are a threat to the external validity of survey research. Measures were obtained on the selected demographic characteristics of the respondents for: age, occupational differences, ethnic and religious differences, education, and housing status.

Studies conducted to assess performance appraisal were reviewed by Landy and Farr (1980). For purposes of the review, they divided their report into sections on: role, context, vehicle, process, and results. Personal characteristics of raters and ratees were investigated by age, gender, ethnic background, and other job-related variables. Bradburn and Sudman (1979) investigated improving interview methods and guestionnaire design. They measured interviewer characteristics for: age, ethnic background, education, and years of experience.

Examples of Age

Interviewer effects for face-to-face interviews was examined by Bradburn and Sudman (1979). They investigated nonprogrammed interviewer behaviors. They used a group of 59 interviewers as subjects. Most interviews were tape recorded (1,049), but some respondents refused to be tape recorded. In addition, there was mechanical failure in some situations. There were 1,172 interviews performed in total. There were 372 questionnaires selected and coded for nonprogrammed interviewer behavior. One-hundred eleven items in each of the 372 questionnaires resulted in frequencies which were based on 41,292 question administrations. Reported below is a modification of their original table for the "Average Number of Speech Behaviors per Question by Interviewer Characteristics." Only the data on age is presented:

Interviewer Characteristics	Reading Errors	Speech Variations	Probes	Feedback	N
Age					
Under 40	.238	.101	.114	.148	16
40-49	.323	.102	.128	.133	18
50 and Over	.307	.136	.165	.190	25

No significant differences were found in the frequencies of behaviors associated with the demographic background characteristics of the interviewers. However, interviewers who were over 50 years of age exhibited higher levels of nonprogrammed behavior. These differences were nonsignificant (Bradburn & Sudman, 1979).

In the Messmer and Seymour (1982) study on the effects of branching on item nonresponse, a Kendall correlational analysis was used to measure the influence of the demographic characteristics of age and education. Out of eight hypotheses, they presented two related to age and education. These hypotheses are presented below:

"The greater the level of education of the respondents, the lesser the frequency of item nonresponse for branching questions."

"The greater the age of the respondent, the greater the frequency of item nonresponse for branching questions."

Messmer and Seymour (1982) correlated these demographic characteristics with the number of errors divided by the number of branches attempted for each respondent. Following is a their table entitled "Correlation Coefficients for Age and Education with Item Nonreponse."

	Kendall Coefficient	N	p-value
Age	0.0453	2,098	.014
Education	-0.0135	2,083	.250

The .014 level of significance was obtained for the frequency of branching nonresponse, and the age of the respondent. Results indicated that there was no significance between the frequency of branching nonresponse as a function of education. Messmer and Seymour (1982) concluded that as the age of the respondent increased, the frequency of item nonresponse for branching items also increased.

Comparisons of Age

Effects of age as a demographic characteristic in questionnaire construction have been measured using a number of different approaches. This can be illustrated by the work of Bradburn and Sudman (1979) where the age of interviewers was investigated. Nonresponse to items, and nonresponse by refusing to participate in surveys, was examined by Messmer and Seymour (1982) and O'Neil (1979). How respondents respond to questionnaire items as a function of demographic characteristics has also been researched regarding age of the respondent (Schuman & Presser, 1981; Landy & Farr, 1980; Bradburn & Sudman, 1979; Lampert, 1979).

Most survey research regarding age as a demographic characteristic focuses on the behavior of the respondent. Face-to-face interviews have not usually been investigated for demographic background of the interviewers. Bradburn and Sudman (1979) investigated the way in which interviewers

ask respondents questions. They analyzed how often nonprogrammed interviewer behaviors occurred for reading errors, errors in recording, speech variations, feedback, methods of probes, and failures to probe. Their findings indicated that about one-half of all item administrations included these nonprogrammed behaviors. Reading errors were the most prevalent. Interviewer characteristics measured were: race (caucasian and black), age (under 40, 40-49, 50 and over), education (no college, some college, graduated from college), and interviewing experience (under 1 year, 1-5 years, over 5 years). The national sample consisted of 1,200 adult respondents and 59 female interviewers. They found no significant differences among frequencies of behavior associated with the demographic characteristics of the interviewer. Interviewers who are 50 years old or over tended to have higher levels of nonprogrammed behavior. This was not statistically significant. They suggested that older interviewers were more casual in their interviewing technique, and they were not as likely to present a standardized survey to the respondent.

Messmer and Seymour (1982) investigated the effect of branching on item nonresponse for a mail survey. They determined that branching instructions significantly increased the rate of item nonresponse at the .0057 level for questions that immediately followed the branch. Older respondents reached a significant level of .014 for item nonresponse, although age for older respondents was never defined. Item nonresponse was not significantly related to other characteristics, such as gender and education.

Most studies for nonresponse rate have been conducted through mail surveys. In an unusual research design, a telephone survey was conducted to measure nonresponse. The adequacy of response rate was investigated using random digit dialing. To increase response rate in a telephone survey, O'Neil (1979) used callbacks of up to 20 calls for individuals who were not at home. Individuals who were nonrespondents on the first call received a persuasive letter and were called back again requesting their inclusion in the survey. Some of the demographic characteristics identified as potentially contributing to survey nonresponse were: age, occupation, family income, education, and race. O'Neil identified individuals who had a proclivity toward being nonrespondents. They were identified by their initial resistance toward being a participant in the survey. Results indicated that those individuals who were resistant to participation in the survey were 65 years or older and caucasian of Polish descent. German and Irish descendents had a lesser propensity than Polish-descent respondents toward survey nonresponse. Subjects who initially were nonrespondents had lower incomes and less education.

Schuman and Presser (1981) conducted studies with formally-balanced items. Opinion questionnaire items were formally balanced by presenting two sides of an issue written in parallel language. Following is an illustration of a formally-balanced question:

> "Some people think the use of marijuana should be made legal. Other people think marijuana use should not be made legal. Which do you favor?"

Effects of age, education, personal information, interest, sex, and race were measured. They were not found to be significant. Background characteristics of respondents is often measured in survey research for the way individuals respond to different survey instruments, and the way they respond to different items. Lampert (1979) developed a new attitude scaling device called the Attitude Pollimeter which is a continuous scale. The Attitude Pollimeter was compared to a verbal scale, a numerical scale, and a bipolar scale. Lampert obtained background characteristic measures on age, sex, and education. This was to determine whether these characteristics would affect the respondent's ability to use the different scales. Education was the only variable that differentiated among subjects. Educational level was significant at the .0001 level. In this particular study, the background characteristics of age and sex apparently did not influence the ability to use the different scales. This study used a random sample selected from a list of eligible voters. It is assumed that the age varied widely, but age categories were not provided in the report.

Research on performance ratings was reviewed by Landy and Farr (1980). They reported findings on demographic characteristics for ages of rater and ratee, and other variables. Performance ratings as a function of ages of rater and ratee for full-time employees, and part-time employees, did not reach levels of significance (reported by Landy & Farr, 1980) in research conducted by Klores (1966) and Bass and Turner (1973). No significant correlation was found between ratings for black full-time employees and age (Bass & Turner, 1973 as cited by Landy & Farr, 1980).

Conclusions Regarding Age

Nonresponse associated with branching and/or surveys has been influenced more by older subjects according to the studies reviewed. However, most survey items do not appear to be influenced by ratings from a particular group. How groups of individuals rate an item probably interacts with more than one demographic characteristic, but not in all circumstances.

When response alternatives are influenced by the age of the respondent, the content of the item may possibly be related to the historical perspective of the different cohort group. The rating of response alternatives by cohort group was illustrated by the work of Bradburn and Sudman (1979). They asked respondents about their use of marijuana and alcohol. They found that the mean age for respondents who had tried marijuana was 29 years. The mean age for respondents who had drunk alcohol in the past year was 41 years old. There appears to be historical-cultural differences for each group of cohorts in our society. This age perspective was reflected toward the use of alcohol and marijuana by these respondents.

Nonresponse to an entire survey, or to specific items in a survey, remains a threat to the validity of the research results. O'Neil (1979) and Messmer and Seymour (1982) designed experiments to focus on nonresponse. They sought to identify background characteristics of respondents which would influence nonresponse behavior. In both studies, age was a variable which influenced nonresponse: nonresponse for participation in a telephone survey, and nonresponse for answering items following branching. Age was the only characteristic which was found to influence item nonresponse following a branch. Nonresponse for survey participation was

related to age, as well as other variables, on a telephone survey. Research on item nonresponse has traditionally been focused on the background characteristics of the respondents, the application of the instrument, and the design of the instrument itself. The research performed by Messmer and Seymour was supported by previous findings. They reported on the work of Ferber (1966) and Craig and McCann (1978), where the age of the respondent was related to nonresponse. Nonresponse behavior increased as age increased above about 60 years.

Age has been established as a characteristic which may influence item nonresponse and survey nonresponse. Research performed on survey nonresponse has been limited. It is difficult to develop experimental designs that measure survey nonresponse. Further research is required in the areas of survey nonresponse and item nonresponse for the background characteristic of age.

Demographic characteristics have been measured for item form and scale format differences. No significant differences were found for age-related responses to formally-balanced items. There were no significant differences for age-related response for the ability to use different scale types (Schuman & Presser, 1981; Lampert, 1979). Bradburn and Sudman (1979) did observe that, in some instances, item content may influence age-related responses by cohorts. Although the research was limited, performance ratings did not appear to be influenced because of the age of the rater or ratee (Landy & Farr, 1980; Klores, 1966; Bass & Turner, 1973).

Most research that takes into account the demographic characteristics of the sample is not psychometrically concerned with the influence of an interviewer when a survey incorporates interviews as part of the survey design. Bradburn and Sudman (1979) determined that older interviewers had more nonprogrammed behavior that younger interviewers. Further research on the nonprogrammed behavior of interviewers would need to be conducted in order to confirm this finding.

CHAPTER VII

QUESTIONNAIRE FORMAT

Questionnaire formats have been compared for a wide variety of physical layouts and different types of scales. This chapter reviews questionnaire formats that have been used in such diverse fields as the military, marketing, and education.

Branching is one aspect of a questionnaire format that can reduce the amount of time it takes to complete a survey. When this type of format is used, it is imperative that the branching instructions be clear. There is the potential for branching to increase item nonresponse for items following a branch. This phenomenon appears to be associated with older (e.g., 60 years) respondents. Branching may also be a useful tool for researchers who believe that their ordering of items has influenced the response distribution. This is a common occurrence where respondents are educated in the topic area by the items themselves. If this is suspected, it would be possible to design a study where there were two questionnaires. They would both have identical items, with the exception that one questionnaire would include branching and the other questionnaire would not. In this way, the stimulus value of the questions could be compared on the two versions of the questionnaire.

Other questionnaire layout variations have focused on the amount of structuring for items, responses, and simulation on forms. In experiments with Navy personnel for a behavioral observation form, it was found that as tasks became more complex, semistructured forms were best. Less complex tasks were best rated using highly-structured formats.

Clarity in the layout of a questionnaire is critical since respondents may inadvertently rate an item which they did not mean to select. This may be even more important for respondents who have a low level of education. Education and preference for scale and format have been found to interact. No one format can be purported to be consistently better than any other format.

Description of Questionnaire Layout

There have been a number of approaches taken by researchers in structuring the physical layout of questionnaires. The layout could consist of items and formats which are structured, semistructured, or unstructured (Nugent, Laabs, & Panell, 1982; Mayer & Piper, 1982; Beltramini, 1982; Bardo & Yeager, 1982), and an orderly sequence of questions (Labaw, 1980). Questionnaire length would be considered a portion of the structuring for physical layout (Mayer & Piper, 1982). Primarily, this section addresses questionnaire layouts that include the comparison of various scales, such as Likert, Behaviorally Anchored Rating Scales (BARS), summated scales, numerical scales, semantic differential scales, and Stapel scales. Vertical and horizontal layouts for these scales are compared.

Examples of Questionnaire Layout

Nugent, Laabs, and Panell (1982) examined three formats used to observe and evaluate behaviors on a performance observation form. The proficiency of the rater at the task being evaluated was also examined. Structured, semistructured, and unstructured performance observation forms were developed to evaluate performance on two types of electronic test equipment (Volt OHM-meter and the oscilloscope). Following are their examples of variation in formats for the behavior observation forms (unstructured, semistructured, and structured).

Example of the Unstructured Observation Form

1. "Was the peak-to-peak amplitude of the signal at Test Point #1 measured properly?" ☐ Passed ☐ Failed

"What errors did you observe?" _____

_____

_____

Example of the Semistructured Observation Form

PROBLEM 1 : AMPLITUDE MEASUREMENT

A. PRELIMINARY ADJUSTMENTS

Intensity/Focus
Input Coupling - AC/DC
Display - Channel A
Probe Connections Correct

MAXIMUM POINTS (4.0)
POINTS ASSIGNED: _____

B. CONTROL SETTINGS

Volts/Division - (.05 - .2 cm)
Time/Division - (1 - 20 sec)
Trigger Level - Stable
Channel A Vernier - CAL

MAXIMUM POINTS (4.0)
POINTS ASSIGNED: _____

C. WAVEFORM ANALYSIS

Amplitude Allowed - (2.5 - 2.8 v)
Amplitude Reported _____

MAXIMUM POINTS (15.0)
POINTS ASSIGNED: _____

D. SAFETY

MAXIMUM POINTS (2.0)
POINTS ASSIGNED: _____

PROBLEM TOTAL _____

PASSED ☐ FAILED ☐

Example of the Structured Observation Form

PROBLEM 1 : AMPLITUDE MEASUREMENT

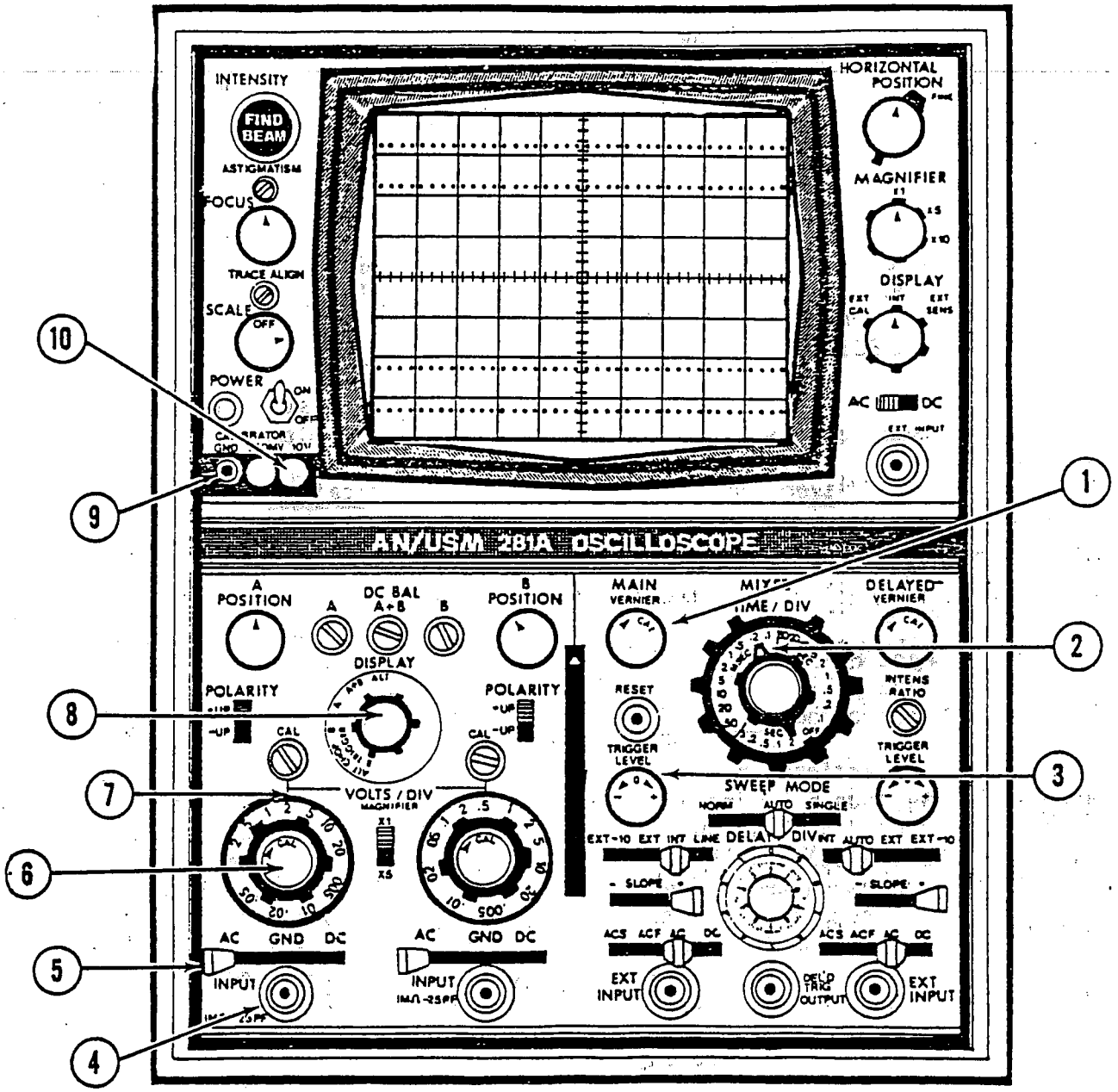
INITIAL SET-UP	PERFORMED CORRECTLY ?	
1. "Was control ⑧ set to the channel A position?"	YES	NO
2. "Was Switch ⑤ set to AC or DC?"	YES	NO
3. "Was the 10:1 probe connected to input jack ④, test point 1, and ground on the black box?"	YES	NO

AMPLITUDE MEASUREMENT PROCEDURE		
1. "Was the final position of Control ⑦ set between .05 and .2 centimeters (cm) deflection?"	YES	NO
2. "Was Control ⑥ set in the CAL position?"	YES	NO
3. "Was a stable waveform displayed (using Control ③ as necessary)?"	YES	NO
4. "Was the number of grid divisions reported between 1.3 and 5.2 centimeters (cm)?"	YES	NO
5. "Was the amplitude of the signal reported between 2.5 and 2.8 volts (v)?"	YES	NO

PASSED ☐ FAILED ☐

Example of the Simulated Observation Form

Adjunct to Structured Observation Form



An initial questionnaire developed and administered by Market Facts of Canada Ltd.'s Consumer Mail Panel required modification due to respondent errors. Respondents mistakenly placed their check marks in wrong categories. Mayer and Piper (1982) provide a before and after example of the modified questionnaire. Originally, respondents meant to mark Brand G, but instead marked the Brand F category by mistake.

The questionnaire layout that confused respondents did not have a response alternative for "other brand." The layout was identical to that of Brand A through Brand G response alternatives (see illustration below). There was no bracketed response alternative for "Other Brand."

"What make or brand is the newest one?"

	Product X	Product Y	Product Z
Brand F ---	( )6	( )6	( )6
Brand G ---	( )7	( )7	( )7
Other brand (SPECIFY)	_____	_____	_____

Mayer and Piper (1982) modified their original questionnaire layout by adding the same response alternative for the category "Other Brand."

Mayer and Piper Format After Modification

"What make or brand is the newest one?"

	Product X	Product Y	Product Z
Brand F ---	( )6	( )6	( )6
Brand G ---	( )7	( )7	( )7
Other brand ___	( )8	( )8	( )8

Various questionnaire layouts have been illustrated in previous sections (see Section 5.1, Response Alternatives; Section 5.3, Number of Scale Points; Section 2.6, Continuous and Circular Scales; and Section 2.1, Multiple-Choice Scales). Beltramini (1982) compared unipolar versus bipolar, number of response alternatives, and horizontal versus vertical scales. In Section 5.1, Response Alternatives, the Stapel scale was exhibited. Essentially, the Stapel scale is a modified and simplified version of the semantic differential scale. Its values range from positive to negative, and measure direction and intensity (Menezes & Elbert, 1979).

Comparisons of Questionnaire Layout

Nugent, Laabs, and Panell (1982) conducted two experiments with Navy personnel. Subjects were instructors and students from the Fleet Anti-Submarine Warfare Training Center. They compared three questionnaire layouts to determine the extent to which the degree of structure influenced rating on observation forms. This was for the operation of electronic test equipment. They were also interested in rater's ability to accurately evaluate performance, as well as rater's own skill level in performing the electronic test equipment task. It was determined that the ability to

perform a task well does not necessarily indicate accurate rating ability. In the first experiment, interrater agreement was as follows: Structured form r = .90, semistructured form r = .58, and unstructured form r = .30. The second experiment indicated different results with interrater agreement: Structured form r = .67, semistructured form r = .72, and unstructured form r = .32. It appears as though a highly structured or semistructured questionnaire layout is superior to a form that is unstructured. Nugent et al. hypothesized that, as a task increases in complexity, a semistructured format may be superior. Highly structured formats may be more appropriate for less complex tasks.

In a marketing study of household members (Market Facts of Canada, Ltd., Consumer Mail Panel), two studies were conducted. The only difference between the two studies was the physical layout of the questionnaire. Mayer and Piper (1982) found layout of the questionnaire to be crucial for self-administered instruments. Their first questionnaire was confusing to respondents. Respondents mistakenly marked the wrong category so that results indicated erroneous brand preferences. Clarity in the physical layout of a questionnaire is essential in obtaining valid results. There is the potential that this type of respondent error can easily go undetected.

Research that compares various combinations of questionnaire layouts has been common. Beltramini (1982) compared variations in scale polarity, number of intervals, and horizontal versus vertical format. Bardo and Yeager (1982) compared number of intervals, Likert and numeric scales, verbal anchors versus numeric anchors at endpoints, and scales anchored with pictures of faces. Borman (1979) compared five formats developed for rating performance in conjunction with a training/no training condition. In a marketing study, Menezes and Elbert (1979) compared Likert, semantic differential, and Stapel scales. Zedeck, Kafry, and Jacobs (1976) examined the degree of agreement on level of rated performance for Behavioral Expectation Scales (BES) for a vertical format, checklist, and graphic rating scale. Bernardin, La Shells, Smith, and Alvares (1976) measured differences in formats for continuous and non-continuous BES.

The investigation of these multiple formats has not, in many instances, supported the superiority of any particular format or scale (Zedeck, Kafry, & Jacobs, 1976; Menezes & Elbert, 1979; Borman, 1979; Beltramini, 1982). The failure to differentiate between questionnaire layouts may be contingent on the quality of the items. Selection of scale items, and item-by-item analysis, are as important as the physical layout of a questionnaire (Beltramini, 1982). Borman (1979) and Zedeck, Kafry, and Jacobs (1976) found that, in comparing various formats and scales to rate performance, no one format was consistently better than another.

Mixed results were obtained by Bernardin, La Shells, Smith, and Alvares (1976) where there was no significant difference found between the ratings in t-tests on the dependent measures (for continuous scales and non-continuous scales). However, separate t-tests on the dependent measures revealed a significant difference for leniency error and discriminant ability between the two formats. They concluded that clarification statements at anchor points had greater rating discriminability and less leniency error at the .05 level of significance for BES.

Bardo and Yeager (1982) observed significant variations among the formats they tested. They found that, regardless of the number of intervals, Likert formats were consistently affected by response set. They defined response set as a psychometric measure where estimates of reliability are inflated, and are a source of systematic error. Systematic error is a potential problem for researchers where respondents consistently use a response set. Respondents tended to rate Likert formats somewhat higher when they were labeled with anchors "strongly agree" to "strongly disagree." There was an indication that increasing the number of scale intervals above five may increase the effects of response set. It was suggested that randomly inverting the order of presentation of item and response alternative may be useful to reduce response set effects.

Another approach to comparing formats was taken by Layne and Thompson (1981). They questioned whether respondents would react to the number of items or to the number of pages in a questionnaire using a Likert-type response format. It was concluded that when 30 items were used, the number of pages in which they were displayed (1 or 3) made no difference. The return rate for this study was only 27.75% even with a follow-up letter. The use of a follow-up letter did not meaningfully increase the response rate.

Conclusions Regarding Questionnaire Layout

Research on questionnaire layout applied to performance evaluation, marketing, and education revealed that no one questionnaire layout was superior to another. Evidence supporting any one layout was sparse and inconsistent. Each questionnaire layout appears to have its own strengths and weaknesses. No one questionnaire layout was consistently better or worse than another.

There is some evidence from the research of Nugent, Laabs, and Panell (1982) and Mayer and Piper (1982) that physical layout and degree of structuring for questions and format may elicit different results. Emphasis needs to be focused on a layout that is clear to respondents. Layout should support respondents in making accurate responses to the intended categories.

7.2 BRANCHING

Description of Branching

In the design of a questionnaire, it may not be feasible or desirable to have every respondent answer all questions. The information requested may not be applicable to all respondents. The approach used to guide respondents through a questionnaire to appropriate questions, but not necessarily all of the questions, has been identified differently by many researchers. Multiple terms used for such identifications have been: branching, leading, routing, filter questions, and screen questions (Messmer & Seymour, 1982; Labaw, 1982; Backstrom & Hurchur-Cesar, 1981).

Screen or filter questions are used to determine how respondents are to be routed through the questionnaire. Some respondents are retained through a sequence of questions, while other respondents move ahead and are eliminated from a set of questions (Backstrom & Hurchur-Ceasar, 1981). Branching requires sets of questions that are integrated instead of questions that would stand alone. Questions are established which will lead the respondent to appropriate subgroups of questions (Labaw, 1982).

Depending on the research design, it is possible to obtain data on branching and nonbranching conditions. One group of respondents receives the branching questionnaire condition, and the second group of respondents receives the questionnaire with no branching. This research design is sometimes used to compare the responses of individuals who are knowledgeable on a subject, and those individuals who are not knowledgeable on a subject (Backstrom & Hurchur-Cesar, 1981). This may give some indication of how much a respondent is learning from the questions themselves, and how the content of a question may be influencing the response to subsequent questions. A second survey could be conducted for those respondents who are found to be knowledgable on a subject. It is possible that the branching questionnaire could be used in lieu of a second questionnaire by filtering and leading some respondents to more in-depth types of items. The responses to branching questionnaires can be compared to the responses to nonbranching questionnaires. The comparison of these two types of questionnaires may assist the researcher in identifying items that require concise wording which is easy to understand. Branching questionnaires should be pretested for clarification and understanding by respondents.

Examples of Branching

Bradburn and Sudman (1979) constructed a questionnaire for a Chicago community study. They measured the main services that the city provided, such as the quality of public schools, library and recreation facilities, police protection, and garbage collection. Following is an example of how they used branching with questions pertaining to voting and transportation:

16. "What part of the day do you usually find most convenient to vote - before 9 a.m., between 9 a.m. and noon, between noon and 5 p.m., or after 5 p.m.?"
    IF NEVER VOTED, SKIP TO Q. 19.

ASK EVERYONE:

19. "Transportation and traffic congestion are two of the major problems of cities today. In general, would you rate Chicago's public transportation system good, fair, or poor?
20. Is the traffic noise where you live loud enough to bother you when you are inside, or is it not a problem?
21. Do you think that the Chicago Police Department does a good, fair, or poor job of controlling traffic?
22. Have you driven a car in Chicago in the last three years? IF NO, SKIP TO Q. 29."

In a draft questionnaire developed by Labaw (1982), branching was used for questions constructed to measure issues related to wills and estates. Respondents were asked whether they had a will. Depending on their answer (yes or no), they were branched to other appropriate questions. The next question at the branch requested information as to the reasons they had for writing a will.

Comparisons of Branching

In research performed by Messmer and Seymour (1982), the effect of branching on item nonresponse was investigated. Initially 4,956 adult subjects received questionnaires, and 2,114 subjects submitted usable questionnaires for analysis. The instrument consisted of 60 items with 10 branching opportunities. Nonresponse rate increased when branching was required. This finding was significant at the .05 level. Education level was not found to influence nonresponses to items associated with branching. They did determine that as the age of respondents increased above approximately 60 years, so did the proportion of item nonresponse. These findings indicate that branching has the potential to increase item nonresponse rates among older respondents.

Conclusions Regarding Branching

Branching is used for questionnaires that are administered through mail, interviews (face-to-face and telephone), and group administration. Researchers need to be careful in their selection of branching. It can be useful for reducing questionnaire completion time and/or interview time. Therefore, branching may be cost effective. Cost effectiveness associated with branching is greatest for questionnaires used in interviews. Branching is not effective in obtaining a 100 percent response rate on all items from group-administered questionnaires or from mailed questionnaires. Questionnaires which incorporate branching, and are mailed out or receive a group administration, may have a shortfall great enough so that investigators need to have a very good reason to employ this technique.

There are alternatives to branching, such as the design of different questionnaire packages for the different categories of respondents. An illustration of this approach was used in the Army Research Institute's test of the Bradley Fighting Vehicle. Four separate questionnaires were designed: one for the driver, one for the track commander, one for the gunner, and one for the remaining personnel.

In situations where the respondents are being interviewed, clear branching instructions are required for the interviewer to make smooth transitions between branches, and to eliminate the potential for a choppy interview. When questionnaires are mailed, branching appears to increase the frequency of nonresponses. This is especially pronounced for older respondents. Items immediately following a branch seem to have an increased rate of nonresponse. This may be due to branching instructions.

CHAPTER VIII

FUTURE RESEARCH

Introduction

This chapter focuses on recommendations for future research which were derived by combining information shortfalls identified from the literature reviews in Chapters II through VII with emerging measurement and computer-based technologies. Background issues in questionnaire research are summarized first to provide a backdrop for the recommendations. Emerging technologies are then summarized to highlight candidate means for improving both the efficiency and effectiveness of questionnaire design and administration for Army Operational Test and Evaluation (OT&E). Higher priority research recommendations are then presented. These are areas where research is expected to produce the most meaningful and timely benefits for Army OT&E. Additional research recommendations are presented in Appendix D.

Background Issues

Results from the experiments reviewed in Chapters I through VII are not in all instances directly applicable for military use without further investigation. For example, even though some of the experiments used military personnel as subjects, the preponderance of experiments used students from universities and colleges. In most instances, there has been a lack of replication across studies. There has also been a lack of consensus as to scale selection, developmental procedures, quantitative analysis, and response characteristics.

One of the reasons that the field of questionnaire construction research has so many inconclusive results is that there has been a paucity of sustained research. Methodological considerations for questionnaire construction require a comprehensive series of experiments. Methodological understanding of questionnaire construction must have continuing research instead of fragmentary research. Tacking questionnaire research on to other studies to investigate occasional methodological issues relegates questionnaire construction issues to a continuing status of inconclusive evidence.

Questionnaire construction research has not progressed evenly across professional fields. In the political arena, social psychologists, political sociologists, and political scientists seek reliable estimates of conceptually valid attitudes in national surveys. To establish demographic and other strong correlates of expressed opinions, great rigor in questionnaire construction is used. Marketing is another area where attitudes, preferences, and perceptions must be reliably estimated by marketing researchers who use computers as a key tool. However, this has not been the case for OT&E.

Emerging Relevant Technologies

In the past few years, computer technologies have had a marked impact on many areas, including information gathering. Previously in questionnaire research, computers were used largely to grade standardized forms, and to collect and analyze experimental data. The role of the computer is changing in the military, as well as in private and other public sectors of society. The impact of computers in transforming questionnaire construction, administration, and scoring is probably attributed primarily to economics. Microprocessor, accessory and software costs have continued to decline (Koenig, 1983; Matarazzo, 1983; Space, 1981). Combining this trend with efficiencies that can result from computer utilization makes the application of computers to questionnaire research, development and application quite attractive.

Computers have brought about many meaningful changes for questionnaire construction in the health sciences. Physicians, psychologists, and psychiatrists now use structured interviews that are performed by computers. Computerized behavioral assessment instruments are being used to screen for problems such as drugs and/or alcohol abuse, and the potential for suicide. Psychological data also are being collected by computer which may be used in diagnosing certain disorders. For example, Space (1981) reported the work of Glaser and Collen (1972) where they selected interview questions using the Bayesian approach (computer adaptive testing) in the diagnosis of diabetes.

Much of the emphasis on computer testing has come from the Navy Personnel Research and Development Center where they have been researching a computerized version of the Armed Services Vocational Aptitude Battery. The Pentagon's plan is to administer these tests at computer terminals, and to expand this computer testing capability so that eventually there may be up to 10,000 computer terminals available for testing by 1986 (Koenig, 1983).

Adaptive testing is being investigated by the armed services (Warm, 1978). The Armed Services Vocational Aptitude Battery is being developed for computer-adaptive testing by the Navy Personnel Research and Development Center (Koenig, 1983). This type of questionnaire design also uses a Bayesian model as a foundation. Each time a question is asked, there is a recalculation of probabilities so that the next item selected is based on the subject's response to the previous item. This allows for estimating the respondent's future performance level as a way to select the next item. The items are administered on a computer, and each respondent receives a different set of questions (Trollip & Anderson, 1982). Computer-adaptive testing has also been known as adaptive testing, tailored testing, stradaptive testing, flexilevel testing, item response theory, characteristic curve theory, and latent trait theory. To the field of questionnaire construction, adaptive testing has probably been the greatest breakthrough in the application of computers so far. Thomas Warm (1978) of the U.S. Coast Guard Institute states that "Item Response Theory (IRT) is the most significant development in psychometrics in many years. It is, perhaps, to psychometrics what Einstein's relativity theory is to physics."

Adaptive testing requires a large sample for its development. Warm (1978) reports that Frederick M. Lord, Educational Testing Service, used a sample size of over 100,000 subjects in 1965. It has been primarily used as an ability test with multiple choice questions. There have been other types of applications such as interviewing subjects for diagnosis of diabetes. The armed forces are a leader in adaptive testing. Even so, currently this model does not appear to be viable for OT&E because of the large samples, and the lead time for development.

This does not mean that there are not many creative uses for computers in OT&E. For example, pilot workload has been assessed using the Subjective Workload Assessment Technique (SWAT) (Shingledecker, 1983). SWAT is based on additive conjoint measurement methodology where ordinal ratings are obtained on variables (time load, mental effort load, and stress load) which are associated with the pilots' subjective feelings of workload. The ordinal ratings are combined into a one-dimensional scale that has interval properties. The ability to derive interval level data from ordinal level data is a major advantage of conjoint measurement. SWAT is being refined and validated for general applicability. This is especially important since the development of subjective measures has usually been situationally specific. In many flight tests or OT&Es, the subjective measures have been selected only for face validity, ease of administration, and minimum intrusiveness. These instruments have not always been accompanied by validity or reliability data (Eggemeier, Crabtree, & La Point, 1983; Eggemeier, McGhee, & Reid, 1983; Eggemeier, Crabtree, Zingg, Reid, & Shingledecker, 1982; Reid, Eggemeier, & Nygren, 1982; Reid, Shingledecker, & Eggemeier, 1981; Reid, Shingledecker, Nygren, & Eggemeier, 1981).

One of the potential advantages of using computers is the savings of time to construct questionnaires. Moroney (1984) has suggested the development of an interactive management information system with expert system capability. To reduce an investigator's time in determining the appropriate specifications and standards in developing questionnaires and checklists, an automated questionnaire generation system could contain a data base capable of generating a questionnaire. Automated systems may be a forerunner of future questionnaire construction because of constraints placed upon investigators for improved efficiency in developing items. This may be due to a stable or decreasing pool of researchers, and an increasing number of systems requiring evaluation (Moroney, 1984). It is cautioned that even automated systems would require pretesting questionnaires, and item reduction for unidemensional and multidimensional scaling.

Artificial intelligence (AI) is another developmental area that may have future value in questionnaire design and use. AI is being applied in industry, government, and defense. Expert systems (ES), as a form of AI, have been created with powerful higher-order languages (HOLs). HOLs excel at symbolic inference (Martins, 1984). Tasks are being identified that are not too complex for ESs (Tate, 1984). Expert systems appear to be effective for relatively simple applications. Knowledge engineers are attempting ES which are less complex, and more practical and realistic than in the past. There have been problems in coding expert systems for real-world application since they are not easy to understand, debug, extend, or maintain. Rule-based paradigms have led to poor computational performance for ES except for the most simplistic application (Martins, 1984).

A typical ES was developed by Teknowledge, Palo Alto, California for application of a structured selection system. A knowledge base is used where there is a finite set of solutions. The actual application was a catalog of equipment. The user could troubleshoot a complex piece of equipment by selecting one of many diagnoses. Texas Instruments, Dallas, Texas is working on a Navy contract to develop future smart weapon systems. This application of ES is expected to be a key to the Strategic Computing Initiative program of the Defense Advanced Research Projects Agency (DARPA) (Verity, 1984).

The term Expert System may imply more capability than might exist. Knowledge engineers are trying to build software solutions to complicated system operations. The operations have been in part non-deterministic, and not closed-end. ES is basically a data base and decision tree combined. Fourth generation computers are capable of integrating several decision trees simultaneously with specialized multiprocessors (Myers, 1984).

Another innovative approach to questionnaires data collection was proposed for survey research with a large subject pool which is geographically dispersed. Surveys have been conducted by using cable television systems to pretest television commercials. The television announcer performs the role of the interviewer, and the respondents are surveyed via telephone (Frankel, 1975). If researchers could apply computers as a feedback device along with the television interviewing, it would be possible to obtain measurements of reactions instantaneously. There may be military application for this type of survey since there are times when large sample sizes are used which may be located in geographically dispersed areas. The Air Force and the Navy have already taken steps to move in the direction of quick response on surveys (but not to the extent mentioned above) through the use of telephone surveys (Chun, Fields, & Friedman, 1975).

The greater incorporation of computers into questionnaire construction has been reviewed for adaptive testing, interviews constructed on a Bayesian model, Subjective Workload Assessment Technique, questionnaire construction systems with software capable of some complex operations, and a combined television-computer technique for large-scale, geographically-dispersed surveys. These are relatively new approaches to the application of computers in survey research. They are all in various developmental stages and require further research for refinement.

Primary Research Recommendations

Priorities have been established among potential research topics as they relate to OT&E performed by the Army Research Institute, Fort Hood, Texas. Priorities are required since the resources available for survey research are limited. Topics were identified which offer the greatest potential for the enhancement of Fort Hood surveys. Eight research recommendations have been highlighted.

These recommendations are associated with: (1) Scale development procedures and analysis, (2) Procedural guides to item wording, (3) Subjective workload assessment methods, (4) Automated Portable Test System, (5) Cognitive complexity, (6) Behaviorally Anchored Rating Scales (BARS), (7) Item nonresponse, branching, and demographic characteristics, and (8)

Pictorial anchors. In addition, other recommendations for future research are ordered according to chapter content, and may be found in Appendix D.

Selection Rationale for Research Recommendations

The eight recommendations selected for future research were identified for their relevance and application to OT&E activities at Fort Hood. They are proposed as research topics because of their potential for meaningful outcomes within a reasonable time frame. There has been a shift in research focus. Previous studies were concerned with variables such as: continuous scales and discrete scales, response alternatives, number of scale points, type of scale format, etc. Because of conflicting research results, it appears as though different scale formats each have their own strengths and weaknesses. More recently, investigation of other variables have focused on: survey developmental procedures, adaptive testing formulated as a computer survey, expert systems, and characteristics of respondents including their cognitive complexity. The eight research recommendations reported in this chapter are not ordered in a priority sequence.

- Scale Development Procedures and Analyses

  Military survey research for the OT&E community needs to investigate ways to obtain more lead time in survey development. Item reduction and multidimensional scaling techniques have been used in commercial-industrial surveys which may be applicable for Army surveys. This would be a vehicle to introduce scale development procedures that would reduce the number of items used in field surveys. For example, Malhotra (1981) designed a developmental procedure that uses different anchors (adjectives, adverbs, and phrases) to measure specific concepts. The scale development procedure includes item reduction, and measures of test reliability and validity. In conjunction with scale development procedures, statistical analyses may benefit by comparing different formulas and statistical assumptions. In comparison of a rank order, paired-comparison, and a Likert scale, data for test-retest reliabilities varied depending on whether a Spearman rho or a Kendall tau was used (Reynolds & Jolly, 1980).

- Procedural Guides to Item Wording

  There is no consensus among survey researchers as to how to word items, and the tone of wording. The influence of wording is not really known. Procedures have been developed to identify specific words that could be used in an item (the Echo technique is an example; see Section 4.2., Wording of Items and Tone of Wording). Various procedures used to identify the use of specific words in an item could be compared because the procedures may possibly identify the structure of the item itself. A method for selecting the item wording requires development to ensure that respondents would only be subjected to items they can understand. Once the method was identified, it may be possible to incorporate the procedures and decision-making processes into an expert system using higher-order languages. Generation of items by an expert system would still require pretesting and possible modification.

- Subjective Workload Assessment Methods

  Assessment of workload is meaningful in OT&E. Continued research is recommended for the general applicability of subjective workload measurement. Subscale analyses used in this method require application to a variety of other types of tasks (Eggemeier, McGhee, & Reid, 1983). Specifically, future research with subjective workload measurement must deal with operational applications of between-subject designs. Common pretraining of subjects, and subjects without common training, may have an effect on between-subject designs. This method has been extensively investigated in the laboratory to demonstrate its validity and reliability. Field applications have been successfully completed in single-place aircraft, multiple-place aircraft, and control room situations. Researchers at Fort Hood could build upon the knowledge and methods gained for programs that measure subjective workload. Subjective workload assessment methods could be used to measure operator workload in Army systems.

- Automated Portable Test Systems

  Administration of surveys could be conducted on a portable test system using a microprocessor which is user-friendly, and contains independent power sources. Entering and collating responses can be performed with accuracy and precision. It is possible to use such a system simultaneously at various remote sites. Information from all locations can be communicated by external cartridge. Questionnaires can be constructed for this type of automated system. Preliminary development of such systems already has been done.

- Cognitive Complexity

  Cognitive compatibility purports to enhance the psychometric quality of ratings when the questionnaire format is compatible with the cognitive structure of the respondent. Cognitive complexity was investigated in an industrial environment, and was shown to be a relevant variable in a rating task. When cognitive complexity was investigated using college/university student samples, there was a failure to replicate previous results. The contextual differences for the type of organization, and the characteristics of the respondents, may have effected the lack of replication. More research is needed to identify military demographic sample characteristics that meaningfully influence questionnaire results.

- Behaviorally Anchored Rating Scales (BARS)

  BARS surveys may be useful in reducing subjectivity found in self-report instruments. This type of scale has been shown to have the capability of replacing self-report measures, and can be used for multiple purposes in addition to the original questionnaire product. It should only be used for large surveys. This may be a useful instrument to develop when multiple applications are required, such as defining objectives, interviewing feedback sessions, and as a foundation for future training programs. BARS could be administered on a portable microprocessor.

- Item Nonresponse, Branching, and Demographic Characteristics

  Branching offers considerable potential for survey efficiency. However, item nonresponse for questionnaires with branches may jeopardize research results. The interaction and/or main effect for item nonresponse, branching, and demographic characteristics for a military sample may be useful in developing new questionnaire formats.

- Pictorial Anchors

  The use of pictorial anchors has been subjected to limited investigation. This methodology could be extended to different types of visually perceived stimuli. It is suggested for possible application with subjects who may have problems with literacy. A variation on the use of pictorial anchors would be the use of color response alternatives, such as the Attitude Pollimeter. This is a color bar in a housing. This nonverbal response alternative could be modified for use with computer graphics so that the respondent could select a gradation in color between two color specturms (see Section 2.6, Continuous and Circular Scales).

BIBLIOGRAPHY

Aiken, L. R. (1978). Reliability, validity and veridicality of questionnaire items. Perceptual and Motor Skills, 47, 161-162.

Aiken, L. R. (1979). Relationships between the item difficulty and discrimination indexes. Educational and Psychological Measurement, 39, 821-824.

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. Educational and Psychological Measurement, 43, 955-959.

Aiken, L. R. (1982). Writing multiple-choice items to measure higher-order educational objectives. Educational and Psychological Measurement, 42, 803-806.

Albaum, G., Best, R., & Hawkins, D. (1981). Continuous vs. discrete semantic differential rating scales. Psychological Reports, 49, 83-86.
Continuous and Circular Scales
Semantic Differential Scales

Allan, P., & Rosenberg, S. (1978, November). Formulating usable objectives for manager performance appraisal. Personnel Journal, 626-642.

Andersen, E. B. (1982, Fall). Latent trait models and ability parameter estimation. Applied Psychological Measurement, 6(4), 445-451.

Andrich, D. (1978, Fall). Scaling attitude items constructed and scored in the Likert tradition. Educational and Psychological Measurement, 38(3), 665-680.

Arima, J. K. (1980, May). Performance vs. paper-and-pencil estimates of cognitive abilities (NPS 54-80-06). Monterey, CA: Naval Post Graduate School. (DTIC No. AD A090614)
Ethnic Background
Gender

Askegaard, L. D., & Umila, B. V. (1982, Fall). An empirical investigation of the applicability of multiple matrix sampling to the method of rank order. Journal of Educational Measurement, 19(3), 193-197.

Atkin, R. S., & Conlon, E. J. (1978, January). Behaviorally anchored rating scales: Some theoretical issues. Academy of Management Review, 119-128.
Behaviorally Anchored Rating Scales

Avery, R. D., & Hayle, J. C. (1974). A Guttman approach to the development of behaviorally based rating scales for systems analysts and programmer/analysts. Journal of Applied Psychology, 59, 61-68.

Backstrom, C. H., & Hurchur-Cesar, G. (1981). Survey research. New York, NY: John Wiley & Sons.
Branching
"Don't Know" Category
Interviewing
Multiple-Choice Scales
Open-Ended Items and Closed-End Items
Paired-Comparison Items
Rank Order Scales
Response Alternatives
Semantic Differential Scales
Wording of Items and Tone of Wording

Barden, R. S. (1980, February). Behaviorally based performance appraisals. The Internal Auditor, 36-43.

Bardo, J. W. (1978). An exact probability test for Likert scales with unequal response probabilities. Southern Journal of Educational Research, 12(3), 181-189.

Bardo, J. W., & Yeager, S. J. (1982). Consistency of response style across types of response formats. Perceptual and Motor Skills, 55, 307-310.
Multiple-Choice Scales
Questionnaire Layout

Barker, D., & Ebel, R. L. (1982). A comparison of difficulty and discrimination values of selected true-false item types. Contemporary Educational Psychology, 7, 35-40.
Balanced Items

Barker, M. S., & Hamovitch, M. (1983, January). Job-oriented basic skills (jobs) program: An evaluation (NPRDC TR 83-5). San Diego, CA: Navy Personnel Research and Development Center. (DTIC No. AD A124150)

Bartlett, T. E., & Linden, L. R. (1974). Evaluating managerial personnel, OMEGA. The International Journal of Management Science, 2(6), 815-819.

Bass, A. R., & Turner, J. N. (1973). Ethnic group differences in relationships among criteria of job performance. Journal of Applied Psychology, 57, 101-109.
Age
Ethnic Background

Bavelas, J. B. (1980). In-house report for professionals and nonprofessionals -- procedural details for the "Echo" technique. Victoria, British Columbia: University of Victoria, Department of Psychology.
Wording of Items and Tone of Wording

Beard, A. D. (1979). Bipolar scales with pictorial anchors: Some characteristics and a method for their use. Applied Psychological Measurement, 3(4), 469-480.
Bipolar Scales

Beatty, R. W., Schneier, C. E., & Beatty, J. R. (1977). An empirical investigation of ratee behavior frequency and ratee behavior change using behavioral expectation scales (BES). Personnel Psychology, 30, 647-657.
Behavioral Expectation Scales

Beaumont, J. G. (1982, October). System requirements for interactive testing. International Journal of Man-Machine Studies, 17(3), 311-320.

Bechtel, G. G. (1980, February). A scaling model for survey monitoring. Evaluation Review, 4(1), 5-41.

Bejar, I. I., & Wingersky, M. S. (1982, Summer). A study of pre-equating based on item response theory. Applied Psychological Measurement, 6(3), 309-325.

Beltramini, R. F. (1982). Rating-scale variations and discriminability. Psychological Reports, 50, 299-302.
Multiple-Choice Scales
Number of Scale Points
Questionnaire Layout
Response Alternatives

Bendig, A. W. (1952a). A statistical report on a revision of the Miami instructor rating sheet. Journal of Educational Psychology, 43, 423-429.
Response Alternatives

Bendig, A. W. (1952b). The use of student rating scales in the evaluation of instructors in introductory psychology. Journal of Educational Psychology, 43, 167-175.
Response Alternatives

Bendig, A. W. (1953). The reliability of self-ratings as a function of the amount of verbal anchoring and the number of categories on the scale. Journal of Applied Psychology, 37, 38-41.
Response Alternatives

Benel, D. C. R., & Benel, R. A. (1976). A further note on sex differences on the semantic differential. British Journal of Social Clinical Psychology, 15, 437-439.
Gender
Semantic Differential Scales

Bernardin, H. J. (1977). Behavioral expectation scales versus summated rating scales: A fairer comparison. Journal of Applied Psychology, 62, 422-427.
Behavioral Expectation Scales

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. Journal of Applied Psychology, 63, 301-308.

Bernardin, H. J., Alvares, K. M., & Cranny, C. J. (1976). A recomparison of behavioral expectation scales to summated scales. Journal of Applied Psychology, 61(5), 564-570.

Bernardin, H. J., & Boetcher, R. (1978, August). The effects of rater training and cognitive complexity on psychometric error in ratings. Paper presented at the meeting of the American Psychological Association, Toronto.
Cognitive Complexity

Bernardin, H. J., Cardy, R. L., & Carlyle, J. J. (1982). Cognitive complexity and appraisal effectiveness: Back to the drawing board? Journal of Applied Psychology, 67(2), 151-160.
Cognitive Complexity

Bernardin, H. J., & Kane, J. S. (1980). A second look at behavioral observation scales. Personnel Psychology, 33, 809-814.
Behavioral Observation Scales

Bernardin, H. J., La Shells, M. B., Smith, P. C., & Alvares, K. M. (1976, February). Behavioral expectation scales: Effects of developmental procedures and formats. Journal of Applied Psychology, 61(1), 75-79.
Behavioral Expectation Scales
Behaviorally Anchored Rating Scales
Questionnaire Layout

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65, 60-66.

Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored rating scales. Journal of Applied Psychology, 66(4), 458-463.
Behaviorally Anchored Rating Scales

Bernardin, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 62(1), 64-69.
Behavioral Expectation Scales
Behaviorally Anchored Rating Scales

Bickley, W. R. (1980, September). Training device effectiveness: Formulation and evaluation of methodology (Report No. RR 1291). U.S. Army Research Institute for the Behavioral and Social Sciences.
Multiple-Choice Scales (AD A122 777)

Bieri, J. (1966). Cognitive complexity and personality development. In O. J. Harvey (Ed.), Experience, structure, and adaptability. New York: Springer.

Biggs, J. B. (1970). Personality correlates of some dimensions of study behavior. Australian Journal of Psychology, 22, 287-297.
"Don't Know" Category

Birnbaum, M. H. (1981, March). Reason to avoid triangular designs in nonmetric scaling. Perception and Psychophysics, 29(3), 291-293.

Bittner, A. C., Carter, R. C., & Krause, M. (1981, November). Performance tests for repeated measures: Moran and computer batteries (NBDL 81R012). New Orleans, LA: U.S. Naval Biodynamics Lab.

Blackburn, R. S. (1982, Spring). Multidimensional scaling and the organizational sciences. Journal of Management, 8(1), 95-103.

Blanz, F., & Gheselli, E. E. (1972). This mixed-standard scale: A new rating system. Personnel Psychology, 25, 185-199.
Cognitive Complexity
Mixed Standard Scales

Block, A. S., & Jouett, M. L. (1978, June). Development of a performance evaluation tool for on-the-job competency appraisal for respiratory therapists -- Subproject III (HRP-0901397). Dallas, TX: American Association for Respiratory Therapy.
Multiple-Choice Scales

Blower, D. J. (1980, July). The bias in the presentation of stimuli when the up and down method is used with forced choice responding (NAMRL-1269). Pensacola, FL: Naval Aerospace Medical Research Laboratory. (DTIC No. AD A089790)

Blower, D. J. (1981, August). Determining visual acuity thresholds: A simulation study of stimulus presentation strategies (NAMRL-1269). Pensacola, FL: Naval Aerospace Medical Research Laboratory, Naval Air Station. (DTIC No. AD A111821)
Multiple-Choice Scales

Boote, A. S. (1981). Reliability testing of psychographic scales. Journal of Advertising Research, 21(5), 53-60.
Number of Scale Points
Response Alternatives

Bordeleau, Y., & Turgeon, B. (1977). Comparison of 3 psychometric methods used in attitude questionnaires. Canadian Journal of Behavioural Science, 9(1), 26-36.

Borman, W. C. (1975). Effects of instruction to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 60, 556-560.
Behaviorally Anchored Rating Scales

Borman, W. C. (1977, December). Consistency of rating accuracy and rating errors in the judgment of human performance. Organizational Behavior and Human Performance, 20(2), 238-252.

Borman, W. C. (1979). Format and training effects on rater accuracy and rater errors. Journal of Applied Psychology, 64, 410-421.
Behaviorally Anchored Rating Scales
Cognitive Complexity
Questionnaire Layout

Borman, W. C. (1981, June). Performance ratings: Comments on the state of the art. In Cecil J. Mullins, AFHRL Conference on Human Appraisal Proceedings. Reviewed and submitted for publication by Lonnie D. Valentin, Jr., Chief, Force Acquisition Branch, Manpower and Personnel Division, Air Force Human Resources Laboratory, AFHRL Technical Paper, 81-20.

Borman, W. C., & Dunnette, M. (1975). Behavior based versus task-oriented performance ratings: An empirical study. Journal of Applied Psychology, 60, 561-565.
Behavioral Expectation Scales
Response Alternatives

Borman, W. C., & Rosse, R. L. (1980, September). Peer ratings: Scoring strategy development and reliability demonstration on Air Force basic trainees (AFHRL-TR-80-28). Minneapolis, MN: Personnel Decisions Research Institute. (DTIC No. AD A090325)

Bradburn, N. M., & Sudman, S. (1979). Improving interview method and questionnaire design (3rd printing). San Francisco, CA: Jossey-Bass Publishers.
Age
Branching
Ethnic Background
Interviewing
Length of Items and Number of Items
Open-Ended Items and Closed-End Items

Bradburn, N. M., Sudman, S., Blair, E., & Stocking, C. (1978, Summer). Question threat and response bias. Public Opinion Quarterly, 42(2), 221-234.

Bradley, R. A. (1982, May). Paired comparisons (FSU Statistics Report No. M615; ONR Technical Report No. 157). Tallahassee, FL: The Florida State University, Department of Statistics. (DTIC No. AD A123877)
Paired-Comparison Items

Brannon, R. (1981). Current methodological issues in paper-and-pencil measuring instruments. Psychology of Women Quarterly, 5(4), 618-627.
Gender
Multiple-Choice Scales

Brown, Curtis A. (1982, May). The effect of factor range on weight and scale values in a linear averaging model. Dissertation Abstracts International, 42(11-B).

Bruvold, W. H. (1977). Reconciliation of apparent nonequivalence among alternative rating methods. Journal of Applied Psychology, 62(1), 111-115.

Burns, A. C., & Harrison, C. (1979). A test of the reliability of psychographics. Journal of Marketing Research, 16, 32-38.

Butler, M. C., & Jones, A. P. (1979). The health opinion survey reconsidered: Dimensionality, reliability, and validity. Journal of Clinical Psychology, 35(3), 554-559.

Carter, R. C., Kennedy, R. S., Bittner, A. C., & Krause, M. (1981, July). Item recognition as a performance evaluation test for environmental research. New Orleans, LA: U.S. Naval Biodynamics Laboratory.

Carter, R. C., & Sbisa, H. E. (1982, January). Human performance tests for repeated measurements: Alternate forms of eight tests by computer (NBDL-82R003). New Orleans, LA: U.S. Naval Biodynamics Lab.

Carter, R. C., Stone, D. A., & Bittner, A. C. (1982). Repeated measurements of manual dexterity applications and support of the two-process theory. Ergonomics, 25(9), 829-838.

Cascio, W. F., & Valenzi, E. R. (1977). Behaviorally anchored rating scores: Effects of education and job experience of raters and ratees. Journal of Applied Psychology, 62, 278-282.
Education

Chapman, R. G., & Staelin, R. (1982, August). Exploiting rank ordered choice set data within the stochastic utility model. Journal of Marketing Research, 19, 288-301.

Checklist: How effective is your management of personnel (1978, September). Focus on Employee Relations, by the Bank Personnel Division staff of the American Bankers Associations.

Christian, J. K., & Bringmann, W. G. (1982). Comparison of computerized versus standardized feedback and accurate versus inaccurate feedback. Psychological Reports, 50, 1067-1070.

Chun, K., Fields, V., & Friedman, S. (1975, August). Military attitudinal surveys: An overview. In H. W. Sinaiko, & L. A. Broedling (Eds.), Perspectives on attitude assessment: Surveys and their alternatives. Manpower Research and Advisory Services, Smithsonian Institution, prepared under the Navy Manpower R&D Program of the Office of Naval Research, N00014.67-A-0399.0006.
Future Research
Interviewing

Church, F. (1983, June). Questionnaire construction manual for operational tests and evaluation. Prepared for the Deputy Commander of Tactics and Test, 57th Fighter Weapons Wing/DT, Tactical Fighter Weapons Center (TFWC), Nellis AFB, NV.
Interviewing
Number of Scale Points
Rank Order Scales
Semantic Differential Scales

Cicchinelli, L. F., Harmon, K. R., & Keller, R. A. (1982, December). Relative cost and training effectiveness of the 6883 F-111 converter/flight control system simulators as compared to actual equipment (AFHRL-TR-82-30). Lowry AFB, CO: Logistics and Technical Training Division.
Multiple-Choice Scales
Open-Ended Items and Closed-End Items

Cocanougher, A. B., & Ivancevich, J. M. (1978, July). "BARS" performance rating for sales force personnel. Journal of Marketing, 87-95.
Behaviorally Anchored Rating Scales

Cole, N. (1973). On measuring the vocational interests of women. Journal of Counseling Psychology, 20, 105-112.
Continuous and Circular Scales

Comrey, A. L., & Montag, I. (1982, Summer). Comparison of factor analytic results with two-choice and seven-choice personality item formats. Applied Psychological Measurement, 6(3), 285-289.
Number of Scale Points

Cooper, W. H. (1981, September). Ubiquitous halo. Psychological Bulletin, 90(2), 218-244.

Cottle, C. E., & McKeown, B. (1980, January). The forced-free distinction in Q technique: A note on unused categories in the Q sort continuum. Operant Subjectivity, 3(2), 58-63.

Couch, A., & Keniston, K. (1960). Yea sayers and nay sayers: Agreeing response set as a personality variable. Journal of Abnormal and Social Psychology, 60, 151-174.
Balanced Items

Craig, C. S., & McCann, J. M. (1978). Item nonresponse in mail surveys: Extent and correlation. Journal of Marketing Research, 15, 285-289.
Age

Cronbach, L. J. (1946). Response sets and test validity. Educational and Psychological Measurement, 6, 475-494.
Balanced Items

Cronbach, L. J. (1950). Further evidence on response sets and test design. Educational and Psychological Measurement, 10, 3-31.
Balanced Items

Crooks, L. A. (Ed.). (1972). An investigation of sources of bias in the prediction of job performance: A six year study. Princeton, NJ: Educational Testing Service.
Ethnic Background

Cudeck, R. (1980, Summer). A comparative study of indices for internal consistency. Journal of Educational Measurement, 17(2), 117-130.

Dambrot, F. (1980, April). Test item order and academic ability, or should you shuffle the test item deck? Teaching of Psychology, 7(2), 94-96.
Order of Items

Daniel, F. R., Jr., & Wagner, E. E. (1982). Differences among Holland types as measured by the hand test: An attempt at construct validation. Educational and Psychological Measurement, 42, 1295-1301.

Daniel, W. W., Schott, B., Atkins, F. C., & Davis, A. (1982, Spring). An adjustment for nonresponse in sample surveys. Educational and Psychological Measurement, 42(1), 57-67.

Deaton, W. L., Glasnapp, D. R., & Poggio, J. P. (1980). Effects of item characteristics on psychometric properties of forced choice scales. Educational and Psychological Measurement, 40, 599-610.
Balanced Items
Multiple-Choice Scales
Wording of Items and Tone of Wording

DeCotiis, T. A. (1977). An analysis of the external validity and applied relevance of three rating formats. Organizational Behavior and Human Performance, 19, 247-266.

DeCotiis, T. A. (1978). A critique and suggested revision of behaviorally anchored rating scales developmental procedures. Educational and Psychological Measurement, 38, 681-690.

De Jung, J. E., & Kaplan, H. (1962). Some differential effects of race of rater and combat attitude. Journal of Applied Psychology, 46, 370-374.
Ethnic Background

Devine, P. J. (1980, November). An investigation of the degree of correspondence among four methods of item bias analysis. Dissertation Abstracts International, 41(5-B).

Dickinson, T. L., & Zellinger, P. M. (1980). A comparison of behaviorally anchored rating and mixed standard scale formats. Journal of Applied Psychology, 65(2), 147-154.
Mixed Standard Scales

Dickson, J., & Albaum, G. (1977). A method for developing tailormade semantic differentials for specific marketing content areas. Journal of Marketing Research, 14, 87-91.
Bipolar Scales
Multiple-Choice Scales
Semantic Differential Scales

Divgi, D. R. (1981). Model-free evaluation of equating and scaling. Applied Psychological Measurement, 5(2), 203-208.

Dixon, D. J., Copeland, M. G., & Halcomb, C. G. (1978, May). Psychomotor battery approaches to performance prediction and evaluation in hyperbaric, thermal and vibratory environments: Annotated bibliographies and integrative review (NBDL-M002). Lubbock, TX: Texas Tech University. (DTIC No. AD A099981)

Dolch, N. A. (1980). Attitude measurement by semantic differential on a bipolar scale. The Journal of Psychology, 105, 151-154.
Bipolar Scales
Response Alternatives
Semantic Differential Scales

Dollard, J. A., Dixon, M. L., & McCann, P. H. (1980, September). Shipboard instruction and training management with computer technology: A pilot application (NPRDC-TR-80-34). San Diego, CA: Navy Personnel Research and Development Center.
Multiple-Choice Scales
Middle Scale Point Position

Downs, P. E. (1978). Miscellany: Testing the upgraded semantic differential. Journal of the Market Research Society, 20(2), 99-103.
Semantic Differential Scales

Downs, S., Farr, R. M., & Colbeck, L. (1978). Self-appraisal: A convergence of selection and guidance. Journal of Occupational Psychology, 51, 271-278.

Drasgow, F. (1982). Biased test items and differential validity. Psychological Bulletin, 92(2), 526-531.

Drasgow, F. (1982, Summer). Choice of test model for appropriateness measurement. Applied Psychological Measurement, 6(3), 297-308.

Drasgow, F., & Miller, H. E. (1982). Psychometric and substantive issues in scale construction and validation. Journal of Applied Psychology, 67(3), 268-279.

Dziuban, C. D., & Shirkey, E. C. (1980). Sampling adequacy and the semantic differential. Psychological Reports, 47, 351-357.
Semantic Differential Scales

Eckblad, G. (1980). The curvex: Simple order structure revealed in ratings of complexity, interestingness, and pleasantness. Scandinavian Journal of Psychology, 21(1), 1-16.

Edvardsson, B. (1980). Effect of reversal of response scales in a questionnaire. Perceptual and Motor Skills, 50, 1125-1126.

Edwards, M. R. (1981, August). Improving performance appraisal by using multiple appraisers. Industrial Management and Data Systems, 13-16.
Paired-Comparison Items

Edwards, R. H. (1981). Coefficients of effective length. Educational and Psychological Measurement, 41, 283-285.

Eggemeier, F. T., Crabtree, M. S., & La Point, P. A. (1983, October). The effect of delayed report on subjective ratings of mental workload. Proceedings of the Human Factors Society 27th Annual Meeting, 139-143.
Future Research

Eggemeier, F. T., Crabtree, M. S., Zingg, J. J., Reid, G. B., & Shingledecker, C. A. (1982). Subjective workload assessment in a memory update task. Proceedings of the Human Factors Society 26th Annual Meeting, 643-647.
Future Research

Eggemeier, F. T., McGhee, J. Z., & Reid, G. B. (1983, May). The effects of variations in task loading on subjective workload rating scales. Proceedings of the IEEE 1983 National Aerospace and Electronics Conference, Dayton, OH, 1099-1105.
Future Research

Eiser, J. R., & Osmon, B. E. (1978). Judgmental perspective and value connotations of response scale labels. Journal of Personality and Social Psychology, 36(5), 491-497.
Balanced Items
Bipolar Scales

Eiser, J. R., & Stroebe, W. (1972). Categorization and social judgment. London: Academic Press.

Eisler, H. (1982). On the nature of subjective scales. Scandinavian Journal of Psychology, 23(3), 161-171.

Elithorn, A., Mornington, S., & Stavron, A. (1982). Automated psychological testing: Some principles and practice. International Journal of Man-Machine Studies, 17(3), 247-263.

Evaluating performance: A self-instruction unit (HRP-0902012), (1979). Princeton, NJ: Educational Testing Service.

Fay, C. H., & Latham, G. P. (1982). Effects of training and rating scales on rating errors. Personnel Psychology, 35, 105-116.
Behavioral Expectation Scales
Behavioral Observation Scales

Ferber, R. (1966). Item nonresponse in a consumer survey. Public Opinion Quarterly, 30, 399-415.
Age

Finley, D. M. (1976, May). The effects of scale continuity and behavioral anchor specificity upon the psychometric properties of performance rating scales. Dissertation Abstracts International, 36(11-B).

Finley, D. M., Osborn, H. G., Dubin, J. A., & Jeanneret, P. R. (1977). Behaviorally based rating scales: Effects of specific anchors and disguised scale continua. Personnel Psychology, 30, 659-669.
Mixed Standard Scales

Finn, R. H. (1976). Concerning questionnaire quality and operational utility. Catalog of Selected Documents in Psychology, 6, 32.

Fischer, G. H., & Formann, A. K. (1982, Fall). Some applications of logistic latent trait models with linear constraints on the parameters. Applied Psychological Measurement, 6(4), 397-416.

Fivars, G. (1975). The critical incident technique: A bibliography. JSAS Catalog of Selected Documents in Psychology, 5, 210.
Behavioral Observation Scales

Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-1358.
Behavioral Observation Scales

Fleishmann, U. (1981). Psychometric techniques and questionnaires for use in gerontopsychological investigations. International Journal of Rehabilitation Research, 4(1), 96-97.

Ford, D. L. (1976, February). Predicting group decision strategies: The effect of rating-scale use bias on accuracy of prediction. Catalog of Selected Documents in Psychology, 6(3).

Fowler, F. J. (1984). Survey research methods. Beverly Hills, CA: Sage Publications.

Fowler, F. J., & Mangione, T. W. (1983). The role of interviewing training and supervision in reducing interviewer effects on survey data. Proceedings of the American Statistical Association Meeting, Survey Research Methods Section, 124-128.

Fralicx, R. D., & Raju, N. S. (1982). A comparison of five methods for combining multiple criteria into a single composite. Educational and Psychological Measurement, 42, 823-827.

Frankel, L. R. (1975, August). Restrictions to survey sampling legal, practical, and ethical. In H. W. Sinaiko, & L. A. Broedling (Eds.), Perspectives on attitude assessment: Surveys and their alternatives. Manpower Research and Advisory Services, Smithsonian Institution, prepared under the Navy Manpower R&D Program of the Office of Naval Research (N00014.67-A-0399.0006).
Future Research

Frederiksen, N., Jensen, O., & Beaton, A. E. (1972). Prediction of organizational behavior. New York: Pergamon Press, Inc.
Semantic Differential Scales

Friedman, B. A., & Cornelius, E. T. (1976). Effect of rater participation in scale construction on the psychometric characteristics of two rating scale formats. Journal of Applied Psychology, 61(2), 210-216.
Behaviorally Anchored Rating Scales

Fullerton, J. T., & Holley, M. (1982). A new look at standard scale transformation. Psychological Reports, 50, 1148-1150.

Furnham, A., & Henderson, M. (1982). The good, the bad and the mad: Response bias in self-report measures. Personality and Individual Differences, 3(3), 311-320.

Garner, W. R. (1960). Rating scales, discriminability, and information transmission. Psychological Review, 67, 343-352.
Number of Scale Points

Gay, L. R. (1980). The comparative effects of multiple-choice versus short-answer tests on retention. Journal of Educational Measurement, 17(1), 45-50.

Gerow, J. R. (1980, April). Performance on achievement tests as a function of the order of item difficulty. Teaching of Psychology, 7(2), 93-94.
Order of Items

Gibbons, J. D., Olkin, I., & Sobel, M. (1979, September). A subset selection technique of scoring items on a multiple choice test. Psychometrika, 44(3), 259-270.

Gividen, G. M. (1973, February). Order of merit: Descriptive phrases for questionnaires. Unpublished report, available from the ARI Field Unit at Fort Hood, TX.
Middle Scale Point Position

Glaser, M. A., & Collen, M. F. (1972). Toward automated medical decisions. Computers and Biomedical Research, 5, 180-189.
Future Research

Goodstadt, M. S., & Magid, S. (1977). When Thurstone and Likert agree -- A confounding of methodologies. Educational and Psychological Measurements, 37, 811-818.

Graef, J., & Spence, I. (1979, January). Using distance information in the design of large multidimensional scaling experiments. Psychological Bulletin, 86(1), 60-66.

Green, B. F. (1981). A primer of testing. American Psychologist, 36(10), 1001-1011.
Multiple-Choice Scales

Groves, R. M. (1979). Actors and questions in telephone and personal interview surveys. Public Opinion Quarterly, 43, 190-205.
Interviewing
Response Alternatives

Guilford, J. P. (1954). Psychometric methods. New York: McGraw-Hill.
Number of Scale Points

Guion, R. M. (1979, April). Principles of work sample testing. II. A non-empirical taxonomy of test uses (TR-79-A8). Bowling Green, OH: Bowling Green State University. (DTIC No. AD A072446)
Number of Scale Points

Guion, R. M., & Ironson, G. H. (1983, February). Latent trait theory for organizational research. Organizational Behavior and Human Performance, 31(1), 54-87.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stauffer, et al. (Eds.), Measurement and prediction. Princeton, NJ: Princeton University Press.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. Lazarfeld (Ed.), Mathematical thinking in the social sciences. Glencoe, IL: Free Press.
Continuous and Circular Scales

Guttman, L. (1957). Empirical verification of the radex structure of mental abilities and personality traits. Educational and Psychological Measurement, 17, 391-407.
Continuous and Circular Scales

Hahn, J. E. (1981, March). A Monte Carlo study of incomplete data designs and configurations in non-metric multidimensional scaling. Dissertation Abstracts International, 41(9-B).

Hambleton, R. K., Mills, C. N., & Simon, R. (1983). Determining the lengths for criterion referenced tests. Journal of Educational Measurement, 20(1), 27-38.

Hambleton, R. K., & van der Linden, W. J. (1982, Fall). Advances in item response theory and applications: An introduction. Applied Psychological Measurement, 6(4), 373-378.

Hamel, C. J., Braby, R., Terrell, W. R., & Thomas, G. (1983, January). Effectiveness of job training materials based on three format models: A field evaluation (Technical Report 138). Orlando, FL: Training Analysis and Evaluation Group, Department of the Navy.
Multiple-Choice Scales

Hamelink, J., & Hamelink, J. (1980). A numeric plan for performance appraisal. Training and Development Journal, 34(10), 88-89.

Hammer, W. C., Kim, J. S., Baird, L., & Bigoness, N. J. (1974). Race and sex as determinants of ratings by potential employers in a simulated work sampling task. Journal of Applied Psychology, 59, 705-711.
Ethnic Background

Hardt, R. H., Eyde, L. D., Primoff, E. S., & Tordy, G. R. (1978). The New York State Trooper job element examination. Albany, NY: New York Police.
Ethnic Background

Hartke, A. R. (1979, Fall). The development of conceptually independent sub-scales in the measurement of attitudes. Educational and Psychological Measurement, 39(3), 585-592.

Harvey, R. J. (1982, April). The future of partial correlation as a means to reduce halo in performance ratings. Journal of Applied Psychology, 67(2), 171-176.

Hatchett, S., & Schuman, H. (1975). White respondents and race-of-interviewer effects. Public Opinion Quarterly, 39, 523-528.
Interviewing

Heavlin, W. D., Lee-Merrow, S. W., & Lewis, V. M. (1982, Fall). The psychometric foundations of goal attainment scaling. Community Mental Health Journal, 18(3), 230-241.

Helm, W. R., & Donnell, M. L. (1979). Mission operability assessment technique: A system evaluation methodology (TP 79-31). Point Mugu, CA: Pacific Missile Test Center. (NTIS No. ADB042746)
Multiple-Choice Scales

Hendel, D. D. (1979). Paired comparisons intransitivity: Is it relatively stable over time? Educational and Psychological Measurement, 39, 779-784.

Hirschman, E. C., & Wallendorf, M. R. (1982). Free-response and card-sort techniques for assessing cognitive content: Two studies concerning their stability, validity, and utility. Perceptual and Motor Skills, 54, 1095-1110.

Holbrook, M. B. (1977). Comparing multiattribute attitude models by optimal scaling. Journal of Consumer Research, 4(3), 165-171.

Holland, P. W. (1981, March). When are item response models consistent with observed data? Psychometrika, 46(1), 79-92.

Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. Journal of Applied Psychology, 63(5), 579-588.

Hom, P. W., De Nisi, A. S., Kinicki, A. J., & Bannister, B. D. (1982). Effectiveness of performance feedback from behaviorally anchored rating scales. Journal of Applied Psychology, 67(5), 568-576.
Behaviorally Anchored Rating Scales

Horayangkura, V. (1978, December). Semantic dimensional structures: A methodological approach. Environment and Behavior, 10(4), 555-584.

Hsu, L. M. (1979). A comparison of three methods of scoring true-false tests. Educational and Psychological Measurement, 39, 785-790.

Hughes, G. D. (1975). Upgrading the semantic differential. Journal of the Marketing Research Society, 17(1), 41-44.
Semantic Differential Scales

Hulin, C. L. (1982, April). Some reflections on general performance dimensions and halo rating error. Journal of Applied Psychology, 67(2), 165-170.

Hulin, C. L., Drasgow, F., & Komocar, J. (1980, July). Applications of item response theory to analysis of attitude scale translations (80-5). Champaign, IL: Department of Psychology, University of Illinois. (DTIC No. AD A087834)

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. Journal of Applied Psychology, 67(6), 818-825.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982, Summer). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. Applied Psychological Measurement, 6(3), 249-260.

Humphreys, M. A. (1982, Winter). Data collection effects on nonmetric multidimensional scaling solutions. Educational and Psychological Measurement, 42(4), 1005-1022.

Hyman, H. H., Cobb, W. J., Feldman, J. J., Hart, C. W., & Stember, C. H. (1954). Interviewing in social research. Chicago: University of Chicago Press.
Ethnic Background
Interviewing

Imada, A. S., & London, M. (1979). Relationship between subjects, scales, and stimuli in research on social perceptions. Perceptual and Motor Skills, 48, 691-697.
Ethnic Background

Innes, J. M. (1977). Extremity and "don't know" sets in questionnaire response. British Journal of Social Clinical Psychology, 16, 9-12.
"Don't Know" Category

Ironson, G. H., & Smith, P. C. (1981, Summer). Anchors away -- the stability of meaning of anchors when their location is changed. Personnel Psychology, 34(2), 249-262.

Ivancevich, J. M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. Journal of Applied Psychology, 64(5), 502-508.
Behavioral Expectation Scales

Ivancevich, J. M. (1980). Behavioral expectation scales versus nonanchored and trait rating systems: A sales personnel application. Applied Psychological Measurement, 4(1), 131-133.
Behavioral Expectation Scales
Response Alternatives

Jacobs, R., Kafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. Personnel Psychology, 33, 595-640.
Behaviorally Anchored Rating Scales
Cognitive Complexity

Jensen, H. E., Massey, I. H., & Valentine, L. D. (1977). Armed Services Vocational Aptitude Battery Development (ASVAB Forms 5, 6 and 7) (Technical Research Note 77-33). Fort Sheridan, IL: U.S. Army Enlisted Processing Command.

Jesteadt, W. (1980). An adaptive procedure for subjective judgments. Perception and Psychophysics, 28, 85-88.
Multiple-Choice Scales

Johnson, J. D. (1981). Effects of the order of presentation of evaluative dimensions for bipolar scales in four societies. The Journal of Social Psychology, 113, 21-27.
Balanced Items
Bipolar Scales

Jones, A. P., Main, D. S., Butler, M. C., & Johnson, L. A. (1982). Narrative job descriptions as potential sources of job analysis ratings. Personnel Psychology, 35, 813-828.

Kafry, D., Zedeck, S., & Jacobs, R. (1976). Short notes: The scalability of behavioral expectation scales as a function of developmental criteria. Journal of Applied Psychology, 61(4), 519-522.
Behavioral Expectation Scales

Kalton, G. (1983). Introduction to survey sampling. Beverly Hills, CA: Sage Publications.

Kane, J. S., & Bernardin, H. J. (1982). Behavioral observation scales and the evaluation of performance appraisal effectiveness. Personnel Psychology, 35, 635-641.
Behavioral Observation Scales

Kane, J. S., & Lawler, E. E. (1979). Performance appraisal effectiveness: Its assessment and determinants. In B. Staw (Ed.), Research in organizational behavior (Vol. 1) (pp. 425-478). Greenwich, CT: JAI Press.

Katcher, B. L., & Bartlett, C. J. (1979, April). Rating errors of inconsistency as a function of dimensionality of behavioral anchors (Research Report No. 84). College Park, MD: University of Maryland, Department of Psychology. (DTIC No. AD A068922)
Mixed Standard Scales

Kearney, W. J. (1976, June). The value of behaviorally based performance appraisals. Business Horizons, 75-83.

Kearney, W. J. (1979, January). Behaviorally anchored rating scales -- MBO's missing ingredient. Personnel Journal, 58(1), 20-25.
Behaviorally Anchored Rating Scales
Multiple-Choice Scales

Keaveny, T. J., & McCann, A. F. (1975). A comparison of behavioral expectation scales and graphic rating scales. Journal of Applied Psychology, 60, 695-703.

Kelly, G. A. (1955). The psychology of personal constructs (Vol. 1). New York: Norton Press.
Cognitive Complexity

Kennedy, R., Bittner, A., Jr., Carter, R., Krause, M., Harbeson, M., McCafferty, D., Pepper, R., & Wiker, S. (1981, July). Performance evaluation tests for environmental research (PETER): Collected papers (NBDL-80R008). New Orleans, LA: Naval Biodynamics Laboratory. (DTIC No. AD A111296)

Kennedy, R. S., Bittner, A. C., Jr., Harbeson, M. M., & Jones, M. B. (1981, November). Perspectives in performance evaluation tests for environmental research (PETER) (NBDL-803004). New Orleans, LA: Naval Biodynamics Laboratory. (DTIC No. AD A111180)

Kennedy, R. S., Jones, M. B., & Harbeson, M. M. (1981, November). Assessing productivity and well-being in Navy workplaces. New Orleans, LA: U.S. Naval Biodynamics Laboratory.

Kesselman, G., Lopez, F. M., & Lopez, F. E. (1982). The development and validation of self-report scored in-basket test in an assessment center setting. Public Personnel Management, 11(3), 228-238.

Kiesler, C. A., Collins, B. E., & Miller, N. (1969). Attitude change: A critical analysis of theoretical approaches. New York: John Wiley & Sons, Inc.
Semantic Differential Scales

Kingstrom, P. O. (1979). The effects of rater-ratee interactions and the format of appraisal interviews on rating characteristics and feedback. Dissertation Abstracts, 39(10-B), 5114.
Interviewing

Kingstrom, P. O., & Bass, A. R. (1981). A critical analysis of studies comparing behaviorally anchored rating scales (BARS) and other rating formats. Personnel Psychology, 34, 263-289.
Behaviorally Anchored Rating Scales

Klimoski, R. J., & London, M. (1974). Role of rater in performance appraisal. Journal of Applied Psychology, 59(4), 445-451.

Klockars, A. J. (1979). Evaluative confounding in the choice of bipolar scales. Psychological Report, 45, 771-775.
Balanced Items
Bipolar Scales
Semantic Differential Scales

Klockars, A. J., King, D. W., & King, L. A. (1981). The dimensionality of bipolar scales in self-description. Applied Psychological Measurement, 5(2), 219-227.
Bipolar Scales
Semantic Differential Scales

Klores, M. S. (1966). Rater bias in forced-distribution ratings. Personnel Psychology, 19, 411-421.
Age

Koch, W. R., & Reckase, M. D. (1978, June). A live tailored testing comparison study of the one- and three-parameter logistic models (Research Report 78-1). Columbia, MO: Department of Educational Psychology, University of Missouri. (DTIC No. AD A058528)

Koenig, R. (1983, April 18). Interest rises in testing by computer. The Wall Street Journal.
Future Research

Korschot, B. C. (1978, July-August). Quantitative evaluation of investment research analysts. Financial Analysts Journal, 41-46.

Kowalski, R. (1984). AI and software engineering. Datamation, 30(18), 92-102.

Krohn, G. S. (1984). LAVM/RV OT II human factors assessment materials. Fort Hood, TX: U.S. Army Research Institute for the Behavioral and Social Sciences (ARI), HQ, TCATA, PERI-SH.
Interviewing
Number of Scale Points

Krus, D. J., & Krus, P. H. (1977, Spring). Normal scaling of the unidimensional dominance matrices: The domain referenced model. Educational and Psychological Measurement, 37(1), 189-193.

Labaw, P. (1980). Advanced questionnaire design (2nd ed.). Cambridge, MA: Abt Books.
Branching
Interviewing
Order of Items
Questionnaire Layout
Wording of Items and Tone of Wording

Lahey, M. A., & Saal, F. E. (1981). Evidence incompatible with a cognitive compatibility theory of rating behavior. Journal of Applied Psychology, 66(6), 706-715.
Cognitive Complexity

Lamont, L. M., & Lundstrom, W. J. (1977, November). Identifying successful industrial salesmen by personality and personal characteristics. Journal of Marketing Research, 14, 517-529.

Lampert, S. I. (1979). The Attitude Pollimeter: A new attitude scaling device. Journal of Marketing Research, 16, 578-582.
Age
Continuous and Circular Scales
Education

Landy, F., & Farr, J. (1980). Performance rating. Psychological Bulletin, 87, 72-107.
Age
Behaviorally Anchored Rating Scales
Education
Ethnic Background
Gender
Response Alternatives

Landy, F. J., & Barnes, J. L. (1979). Scaling behavioral anchors. Applied Psychological Measurement, 3(2), 193-200.
Behaviorally Anchored Rating Scales
Paired-Comparison Scales

Landy, F. J., Barnes-Farrell, J. L., Vance, R. J., & Steele, J. W. (1980). Statistical control of halo error in performance ratings. Journal of Applied Psychology, 65(5), 501-506.

Landy, F. J., Farr, J. L., Saal, F. E., & Freytag, W. R. (1976). Behaviorally anchored scales for rating the performance of police officers. Journal of Applied Psychology, 61(6), 750-758.
Behaviorally Anchored Rating Scales

Landy, F. J., Vance, R. J., & Barnes-Farrell, J. L. (1982, April). Statistical control of halo: A response. Journal of Applied Psychology, 67(2), 177-180.

La Rocco, J. M., & Butler, M. C. (1977, December). Survey questionnaires: More than meets the eye (77-57). San Diego, CA: Naval Health Research Center. (DTIC No. AD A100262)

Latham, G. P., Fay, C. H., & Saari, L. M. (1979). The development of behavioral observation scales for appraising the performance of foremen. Personnel Psychology, 32, 299-311.
Behavioral Expectation Scales
Behavioral Observation Scales

Latham, G. P., Saari, L. M., & Fay, C. (1980). BOS, BES, and baloney: Raising kane with Bernardin. Personnel Psychology, 33, 815-821.

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. Personnel Psychology, 30, 255-268.
Behavioral Observation Scales

Latham, G. P., & Wexley, K. N. (1981). Increasing productivity through performance appraisal. Reading, MA: Addison-Wesley.

Layne, B. H., & Thompson, D. N. (1981). Questionnaire page length and return rate. The Journal of Social Psychology, 113, 291-292.
Length of Items and Number of Items
Questionnaire Layout

Lee, R., Malone, M., & Greco, S. (1981). Multitrait-multimethod-multirater analysis of performance ratings for law enforcement personnel. Journal of Applied Psychology, 66(5), 625-632.

Lee, R., Miller, K. J., & Graham, W. K. (1982). Corrections for restriction of range and attenuation in criterion-related validation studies. Journal of Applied Psychology, 67(5), 637-639.

Levine, E. L., Flory, A., & Ash, R. A. (1977). Self-assignment in personnel selection. Journal of Applied Psychology, 62(4), 428-435.
Ethnic Background

Lienert, G. A., & Raatz, U. (1981, Spring). Item homogeneity defined by multivariate symmetry. Applied Psychological Measurement, 5(2), 263-269.

Link, S. W. (1982, September). Correcting response measures for guessing and partial information. Psychological Bulletin, 92(2), 469-486.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. Journal of Applied Psychology, 60, 10-13.
Number of Scale Points

Locander, W. B., & Staples, W. A. (1978). Evaluating and motivating salesmen with the BARS method. Industrial Marketing Management, 7, 43-48.
Behaviorally Anchored Rating Scales

Lodge, M. (1981). Magnitude scaling: Quantitative measurement of opinions. Sage University Paper Series: Quantitative Applications in the Social Sciences, 25.

London, M., & Poplowski, J. R. (1976). Effects of information on stereotypic development in performance appraisal and interview contexts. Journal of Applied Psychology, 61(2), 199-205.
Gender

Lord, F. M. (1982, Fall). Standard error of an equating by item response theory. Applied Psychological Measurement, 6(4), 463-472.

Lund, T. (1975, April). An alternative content method for multidimensional scaling. Multivariate Behavioral Research, 10(2), 181-191.

Mabe, P. A., III, & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. Journal of Applied Psychology, 67(3), 280-296.

Malhotra, N. A. (1981). A scale to measure self-concepts, person concepts, and product concepts. Journal of Marketing Research, 15, 456-464.
Future Research
Semantic Differential Scales

Mann, I. T., Phillips, J. L., & Thompson, E. G. (1979). An examination of methodological issues relevant to the use and interpretation of the semantic differential. Applied Psychological Measurement, 3, 213-229.
Bipolar Scales
Semantic Differential Scales

Marascuilo, L. A., & Slaughter, R. E. (1981, Winter). Statistical procedures for identifying possible sources of item bias based on -sub (x) -sup-2 statistics. Journal of Educational Measurement, 18(4), 229-248.

Marshall, S. P. (1981). Sequential item selection: Optimal and heuristic policies. Journal of Mathematical Psychology, 23, 134-152.

Martin, W. S. (1978, May). Effects of scaling on the correlation coefficient: Additional considerations. Journal of Marketing Research, 15, 304-308.

Martins, G. R. (1984). The overselling of expert systems. Datamation, 30(18), 76-80.
Future Research

Massey, R. H., Mullings, C. J., & Earles, J. A. (1978, December). Performance appraisal ratings: The content issue (AFHRL TR-78-69). Brooks Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory. (DTIC No. AD A064690)

Matarazzo, J. D. (1983). Computerized psychological testing. Science, 221, (4608).
Future Research

Mathews, J. L., Wright, C. E., Yudowitch, K. L., Geddie, J. C., & Palmer, R. L. (1978, August). The perceived favorableness of selected scale anchors and response alternatives (Technical Paper 319). Palo Alto, CA: Operations Research Associates, and Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (DTIC No. AD A061755)
Middle Scale Point Position
Response Alternatives

Maul, T. L., & Pargman, D. (1975). The semantic differential as a psychometric instrument for behavioral research in sport. International Journal of Sport Psychology, 9(1), 7-15.
Semantic Differential Scales

Maurelli, V. A., & Weiss, D. J. (1981, November). Factors influencing the psychometric charracteristics of an adaptive testing strategy for test batteries (Research Report 81-4). Minneapolis, MN: Department of Psychology, University of Minnesota. (DTIC No. AD A109666)

Mayer, C. S., & Piper, C. (1982). A note on the importance of layout in self-administered questionnaires. Journal of Marketing Research, 19(3), 390-391.
Questionnaire Layout

Mayerberg, C. K., & Bean, A. G. (1978, Fall). Two types of factors in the analysis of semantic differential attitude data. Applied Psychological Measurement, 2(4), 469-480.

McBride, J. R., Sympson, J. B., Vale, C. D., Pine, S. M., & Bejar, I. I. (1977, March). Applications of computerized adaptive testing (Research Report 77-1). Minneapolis, MN: Department of Psychology, University of Minnesota. (DTIC No. AD A038114)

McCormick, C. C., & Kavanagh, J. A. (1981). Scaling interpersonal checklist items to a circular model. Applied Psychological Measurement, 5(4), 421-447.
Continuous and Circular Scales

McDonald, R. P. (1981). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 34(1), 100-117.

McDonald, R. P. (1982, Fall). Linear versus nonlinear models in item response theory. Applied Psychological Measurement, 6(4), 379-396.

McFarland, S. G. (1981). Effects of question order on survey responses. Public Opinion Quarterly, 45, 208-215.
Gender
Order of Items

McIver, J. P., & Carmines, E. G. (1981). Unidimensional scaling. Sage University Paper series on quantitative applications in the social sciences, 07-024. Beverly Hills and London: Sage Publishers.
Behavioral Expectation Scales
Behavioral Observation Scales
Mixed Standard Scales
Multiple-Choice Scales
Paired-Comparison Items
Rank Order Scales

McKelvie, S. J. (1978). Graphic rating scales -- how many categories? British Journal of Psychology, 69, 185-202.
Continuous and Circular Scales
Number of Scale Points
Semantic Differential Scales

Meisels, M., & Ford, L. H. (1969). Social desirability response set and semantic differential judgments. Journal of Social Psychology, 78, 45-54.
Gender

Mellenbergh, G. J. (1982, Summer). Contingency table models for assessing item bias. Journal of Educational Statistics, 7(2), 105-118.

Melzer, C. W., Koeslag, J. H., & Schach, S. S. (1981). Correction of item-test correlations and attempts at improving reproducibility in item-analysis: An experimental approach. Educational and Psychological Measurement, 41(4), 979-990.

Menezes, D., & Elbert, N. F. (1979). Alternative semantic scaling formats for measuring store image: An evaluation. Journal of Marketing Research, 16(1), 80-87.
Questionnaire Layout
Response Alternatives

Meriwether, T. N. (1979, May). Developing a job analysis based performance appraisal system at the United States Military Academy: A new approach to forced-choice evaluation. Dissertation Abstracts International, 39(11-B).

Messick, D. M., & Van de Geer, J. P. (1981, November). A reversal paradox. Psychological Bulletin, 90(3), 582-593.

Messmer, D. J., & Seymour, D. T. (1982, Summer). The effects of branching in item response. Public Opinion Quarterly, 46(2), 270-277.
Age
Branching
Education

Meyer, H. H. (1980). Self-appraisal of job performance. Personnel Psychology, 33, 291-295.

Miller, P. M. (1974). A note in sex differences on the sementic differentials. British Journal of Social Clinical Psychology, 13, 33-36.
Gender

Mokken, R. J., & Lewis, C. (1982, Fall). A nonparametric approach to the analysis of dichotomous item responses. Applied Psychological Measurement, 6(4), 417-430.

Mooney, G. R. (1981, September). An analysis of rater error in relation to ratee job characteristics. Dissertation Abstracts International, 42(3-B).

Moroney, W. F. (1984). The use of checklists and questionnaires during system and equipment test and evaluation. Shrivenham, England: NATO Defense Research Group Panel VIII Workshop, Applications of Systems Ergonomics to Weapon System Development, Royal Military College of Science, Vol 1, C-59-C-68.
Future Research
Multiple-Choice Scales

Motowidlo, S. J., & Borman, W. C. (1977). Behaviorally anchored scales for measuring morale in military units. Journal of Applied Psychology, 62(2), 177-183.
Behaviorally Anchored Rating Scales

Mueller, J. E. (1973). War, presidents and public opinion. New York: John Wiley & Sons.
Wording of Items and Tone of Wording

Mullins, C. J. (1981, June). AFHRL conference on human appraisal: Proceedings (AFHRL-TP-81-20). Brooks Air Force Base, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. (DTIC No. AD A102755)

Mullins, C. J., Earles, J. A., & Wilbourn, J. M. (1979, May). Personnel rating effectiveness as a function of number of rating statements (AFHRL-TR-79-11). Brooks Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory.
Length of Items and Number of Items

Munson, J. M., & McIntyre, S. H. (1979). Developing practical procedures for the measurement of personal values in cross-cultural marketing. Journal of Marketing Research, 16, 48-52.
Paired-Comparison Scales

Murphy, J. W. (1980). Use of behaviorally anchored rating scales (BARS) to complement the management by objectives (MBO) and fitness report components of the Marine Corps performance evaluation system. Master of Military Art and Sciences (MMAS) thesis prepared at U.S. Army Command and General Staff College, Fort Leavenworth, KS. (DTIC No. AD A097694)
Behaviorally Anchored Rating Scales

Murphy, K., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. (1982). Relationship between observational accuracy and accuracy in evaluating performance. Journal of Applied Psychology, 67, 320-325.

Murphy, K. R. (1982, April). Difficulties in the statistical control of halo. Journal of Applied Psychology, 67(2), 161-164.

Murphy, K. R., Martin, C., & Garcia, M. (1982). Do behavioral observation scales measure observation? Journal of Applied Psychology, 67(5), 562-567.
Behavioral Observation Scales

Myczyk, J. P., & Gable, M. (1981). Unidimensional (global) vs. multidimensional composite performance appraisals of store managers. Journal of Academic Marketing Science, 9(3), 191-205.

Myers, E. (1984). Business takes the fifth. Datamation, 30(195), 53-57.
Future Research

Neill, J. A., & Jackson, D. N. (1976, Spring). Minimum redundancy item analysis. Educational and Psychological Measurement, 36(1), 123-134.

Nemeroff, W. F., & Wexley, K. N. (1979). An exploration of the relationships between performance feedback interview characteristics and interview outcomes as perceived by managers and subordinates. Journal of Occupational Psychology, 52. 25-34.
Interviewing

Nevo, B. (1980, Summer). Item analysis with small samples. Applied Psychological Measurement, 4(3), 323-329.

Ng, S. H. (1982). Choosing between the ranking and rating procedures for the comparison of values across cultures. European Journal of Social Psychology, 12(2), 169-172.

Norton, S. D., Gustafson, D. P., & Foster, C. E. (1977). Assessment for management potential: Scale design and development, training effects and rater/ratee sex effects. Academy of Management Journal, 20, 117-131.
Gender

Nugent, W. A., Laabs, G. J., & Panell, R. C. (1982, February). Performance test objectivity: A comparison of rater accuracy and reliability using three observation forms (NPRDC TR 82-30). San Diego, CA: Navy Personnel Research and Development Center. (DTIC No. AD A111077)
Questionnaire Layout

Null, C. H. (1980, April). Design considerations for multidimensional scaling. Behavior Research Methods and Instrumentation, 12(2), 274-280.

Nunnally, J. C. (1967). Psychometric theory. New York: McGraw-Hill.
Number of Scale Points

Nunnally, J. C., Leonard, L. C., & Wilson, W. H. (1977). Studies of voluntary visual attention -- theory, methods, and psychometric issues. Applied Psychological Measurement, 1, 203-218.

Nygren, T. E. (1982, April). Conjoint measurement and conjoint scaling: A users guide (AFAMRL-TR-82-22). Wright-Patterson Air Force Base, OH: Air Force Aerospace Medical Research Laboratory, Aerospace Medical Division.

Oaster, T. R. F. (1982). Index for differentiation of state and trait scales with new stressor scales. Psychological Reports, 51, 272.

Oborne, D. J. (1976). Examples of the use of rating scales in ergonomics research. Applied Ergonomics, 7(4), 201-204.
Continuous and Circular Scales

O'Neil, M. J. (1979). Estimating the non-response bias due to refusals in telephone surveys. Public Opinion Quarterly, 43, 218-232.
Age
Education
Ethnic Background

Orlich, D. C. (1978). Designing sensible surveys. Pleasantville, NY: Redgrave Publishing Company.
Behavioral Observation Scales
Interviewing
Multiple-Choice Scales
Open-Ended Items and Closed-End Items
Rank Order Scales
Wording of Items and Tone of Wording

Orpen, C. (1981). The effect of examiner ethnicity on the job satisfaction responses of Blacks in community surveys: A South African study. Journal of Community Psychology, 9(1), 81-85.

Ory, J. C. (1982). Item placement and wording effects on overall ratings. Educational and Psychological Measurement, 42, 767-775.
Balanced Items
Wording of Items and Tone of Wording

Osgood, C. E., Suci, G., & Tannenbaum, P. (1957). The measurement of meaning. Urbana, IL: University of Illinois Press.
Bipolar Scales
Gender
Semantic Differential Scales

Palachek, A. D., & Kerin, R. A. (1980, August). Alternative approaches to the two-group concordance problem in brand preference rankings. Journal of Marketing Research, 386-389.

Paradise, L. V., & Kottler, J. (1979, August). Use of Q-factor analysis for initial instrument validation. Psychological Reports, 45(1), 139-143.

Parkinson, R. (1977, November). Recipe for a realistic appraisal system. Personnel Management, 37-40.

Parsons, C. K., & Hulin, C. L. (1982). An empirical comparison of item response theory and hierarchical factor analysis in applications to the measurement of job satisfaction. Journal of Applied Psychology, 67(6), 826-834.

Peabody, D. (1967). Trait inferences: Evaluative and descriptive aspects. Journal of Personality and Social Psychology Monograph, 7, pp. 644.

Peterson, J. M. (1977, August). The influence of favorable context on questionnaire response. Dissertation Abstracts International, 381(2-A), 747.

Petz, B., & Mayer, D. (1977). Comparison between Thurstone's Law of Comparative Judgments scale and the Sum of Totals scale. Acta Instituti Psychologici, No. 79-86, 55-64.

Presser, S., & Schuman, H. (1980). The measurement of a middle position in attitude surveys. Public Opinion Quarterly, 44(1), 70-85.
Education
Middle Scale Point Position

Primoff, E. S. (1978, August). The use of self-assessments in examining. A paper presented at the 86th Annual Convention of the American Psychological Association, Toronto, Ontario, Canada. (NTIS No. PB-298 606)

Pulakos, E. D. (1984, May). The development of training programs to increase accuracy with different rating formats (84-2). East Lansing, MI: Department of Psychology, Michigan State University.

Pursell, E. D., Campion, M. A., & Gaylord, S. R. (1980, November). Structured interviewing: Avoiding selection problems. Personnel Journal, 907-912.

Pursell, E. D., Dossett, D. L., & Latham, G. P. (1980). Obtaining valid predictors by minimizing rating errors in the criterion. Personnel Psychology, 33, 91-97.

Range, L. M., Anderson, H. N., & Wesley, A. L. (1982, October). Personality correlates of multiple choice answer-changing patterns. Psychological Reports, 51(2), 523-527.

Ray, J. J. (1980). The comparative validity of Likert, projective, and forced-choice indices of achievement motivation. The Journal of Social Psychology, 111, 63-72.
Multiple-Choice Scales

Ray, J. J. (1982). The construct validity of balanced Likert scales. The Journal of Social Psychology, 118, 141-142.
Balanced Items

Ray, J. J., & Bozek, R. S. (1979, June). NSCALE II: A program to analyze and score multiscale surveys and test batteries. Behavior Research Methods and Instrumentation, 11(3), 402.

Reid, G. B., Eggemeier, F. T., & Nygren, T. E. (1982). An individual differences approach to SWAT scale development. Proceedings of the Human Factors Society 26th Annual Meeting, 639-642.
Future Research

Reid, G. B., Shingledecker, C. A., & Eggemeier, F. T. (1981). Application of conjoint measurement to workload scale development. Proceedings of the Human Factors Society 25th Annual Meeting, 522-526.
Future Research

Reid, G. B., Shingledecker, C. A., Nygren, T. E., & Eggemeier, F. T. (1981, October). Development of multidimensional subjective measures of workload. Proceedings of the International Conference on Cybernetics and Society, sponsored by IEEE Systems, Man and Cybernetics Society, Atlanta, GA, 403-406.
Future Research

Resnick, S. M., & Mohrman, A. M. (1982, October). The design of performance appraisal systems: Some implication from research findings (G 8205-924). Los Angeles, CA: University of Southern California, Graduate School of Business Administration. (DTIC No. AD A120333)

Reynolds, T. J. (1976, June). The analysis of dominance matrices: Extraction of unidimensional orders within a multidimensional context (Technical Report No. 3). Los Angeles, CA: Department of Psychology, University of Southern California. (DTIC No. AD A029450)

Reynolds, T. J. (1981, Fall). ERGO: A new approach to multidimensional item analysis. Educational and Psychological Measurement, 41(3), 643-659.

Reynolds, T. J., & Jolly, J. P. (1980). Measuring personal values: An evaluation of alternative methods. Journal of Marketing Research, 17, 531-536.
Future Research
Paired-Comparison Items
Rank Order Scales
Response Alternatives

Ridgway, J., MacCullough, M. J., & Mills, H. E. (1982). Some experiences in administering a psychometric test with a light pen and microcomputer. International Journal of Man-Machine Studies, 17(3), 265-278.

Rigney, J. W., Towne, D. M., Moran, P. J., & Mishler, R. A. (1980, July). Field evaluation of the generalized maintenance trainer-simulator, II. AN/SPA-66 Radar Repeater (NPRDC TR-8-30-2). San Diego, CA: Navy Personnel Research and Development Center.
Rank Order Scales

Rizzo, W. A., & Frank, F. D. (1977, Fall). Influence of irrelevant cues and alternate forms of graphic rating scales on the halo effect. Personnel Psychology, 30(3), 405-417.

Robertson, I. T., & Kandola, R. S. (1982). Work sample tests: Validity, adverse impact and applicant reaction. Journal of Occupational Psychology, 55, 171-183.

Rokeach, M. (1973). The nature of values. New York: Free Press.
Paired-Comparison Items

Roscoe, J. T. (1975). Fundamental research statistics for the behavioral sciences. New York: Holt, Rinehart, and Winston, Inc.
Multiple-Choice Scales

Rose, G. L. (1978). Sex effects on effort attributions in managerial performance evaluation. Organizational Behavior and Human Performance, 21, 367-378.
Gender

Rosinger, G., Myers, L. B., Levy, G., Loar, M., Mohrman, S. A., & Stock, J. R. (1982). Development of behaviorally based performance appraisal system. Personnel Psychology, 35, 75-88.
Mixed Standard Scales

Rounds, J. B., Jr., Miller, T. W., & Dawis, R. V. (1978, Summer). Comparability of multiple rank order and paired comparison methods. Applied Psychological Measurement, 2(3), 415-422.

Roy, J. J. (1982). Machieavellianism, forced-choice formats and the validity of the F scale: A rejoinder to Bloom. Journal of Clinical Psychology, 38(4), 779-782.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980, Spring). A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 17(1), 1-10.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980, Fall). Biased item detection techniques. Journal of Educational Statistics, 5(2), 213-233.

Ryans, A. B., & Srinivasan, V. (1979, November). Improved method for comparing rank-order preferences of two groups of consumers. Journal of Marketing Research, 16, 583-587.

Saal, F. E. (1979). Mixed standard rating scale: A consistent system for numerically coding inconsistent response combinations. Journal of Applied Psychology, 64(4), 422-428.
Cognitive Complexity
Mixed Standard Scales

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88(2), 413-428.

Saito, T. (1977, July). Multidimensional Thurstonian scaling with an application to color metrics. Japanese Psychological Research, 19(2), 78-89.

Samejima, F. (1979, February). Constant information model: A new, promising item characteristic function (Research Report 79-1). Knoxville, TN: Department of Psychology, University of Tennessee. (DTIC No. AD A070090)

Samejima, F. (1979, December). A new family of models for the multiple-choice item (Research Report 79-4). Knoxville, TN: Department of Psychology, University of Tennessee. (DTIC No. AD A080350)

Sauser, W. I. (1979). A comparative evaluation of the effects of rater participation and rater training on characteristics of employee performance appraisal ratings and related mediating variables. Dissertation Abstracts International, 39 (10-B), 5116.

Sauser, W. I., & Pond, S. B. (1981). Effects of rater training and participation of cognitive complexity: An exploration of Schneier's cognitive reinterpretation. Personnel Psychology, 34, 563-577.
Cognitive Complexity

Schaefer, B. A., Bavelas, J., & Bavelas, A. (1980). Using echo technique to construct student-generated faculty evaluation questionnaires. Teaching of Psychology, 7(2), 83-86.
Wording of Items and Tone of Wording

Schaefer, E. S. (1961). Converging conceptual models for maternal behavior and for child behavior. In J. Glidewell (Ed.), Parental attitudes and child behavior. Springfield, IL: Thomas.
Continuous and Circular Scales

Schaeffer, N. (1980). Evaluating race-of-interviewer effects in a national survey. Sociological Methods and Research, 8, 400-419.
Ethnic Background
Interviewing

Schein, V. E. (1973). The relationship between sex role stereotypes and requisite management characteristics. Journal of Applied Psychology, 57, 95-100.
Gender

Schertzer, C. B. (1982). Semantic properties of commonly used scaling adjectives. Dissertation Abstracts International, 42(8-A), 3733.

Schmidt, F. L., & Hunter, J. E. (1980, September). Differential and simple-group validity of employment tests by race: A critical analysis of three recent studies. Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development Center, Research Branch.
Ethnic Background

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1982). Progress in validity generalization: Comments on Callender and Osburn and further developments. Journal of Applied Psychology, 67(6), 835-845.

Schmidt, F. L., & Johnson, R. H. (1973). Effect of race on peer ratings in an industrial setting. Journal of Applied Psychology, 57, 237-241.
Ethnic Background

Schmitt, J. C., & Scheirer, C. J. (1977). The effect of item order on objective tests. Teaching of Psychology, 4, 144-145.
Order of Items

Schneier, C. E. (1977a). Operational utility and psychometric characteristics of behavioral expectation scales: A cognitive reinterpretation. Journal of Applied Psychology, 62(5), 541-548.
Behavioral Expectation Scales
Behaviorally Anchored Rating Scales
Cognitive Complexity
Mixed Standard Scales

Schneier, C. E. (1977b). Multiple rater groups and performance appraisal. Public Personnel Management, 6(1), 13-20.

Schneier, C. E., & Beatty, R. W. (1979a). Integrating behaviorally-based and effectiveness-based methods. The Personnel Administrator, 65-76.
Behaviorally Anchored Rating Scales

Schneier, C. E., & Beatty, R. W. (1979b). Developing behaviorally anchored rating scales. The Personnel Administrator, 59-68.
Behaviorally Anchored Rating Scales

Schneier, C. E., & Beatty, R. W. (1979c). Combining BARS and MBO: Using an appraisal system to diagnose performance problems. The Personnel Administrator, 51-60.
Behaviorally Anchored Rating Scales

Schonemann, P. H. (1982, June). A metric for bounded response scales. Bulletin of the Psychonomic Society, 19(6), 317-319.

Schriesheim, C. A. (1981). The effect of grouping or randomized items on leniency response bias. Educational and Psychological Measurement, 41, 401-411.

Schriesheim, C. A. (1981). Leniency effects on convergent and discriminant validity for grouped questionnaire items: A further investigation. Educational and Psychological Measurement, 41, 1093-1099.

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. Educational and Psychological Measurement, 41, 1101-1114.

Schroder, H. M. (1971). Conceptual complexity and personality organization. In H. M. Schroder & P. Suedfeld (Eds.), Personality theory and information processing. New York: Ronald.

Schuman, H., & Converse, J. (1971). The effects of black and white interviewers in black responses in 1968. Public Opinion Quarterly, 35, 44-68.
Ethnic Background
Interviewing

Schuman, H., & Presser, S. (1981). Questions and answers in attitude surveys: Experiments on question form, wording, and context. New York: Academic Press, Inc.
Age
Balanced Items
"Don't Know" Category
Education
Ethnic Background
Gender
Middle Scale Point Position
Multiple-Choice Scales
Open-Ended Items and Closed-End Items
Order of Items
Wording of Items and Tone of Wording

Schutz, H. G., & Rucker, M. H. (1975). A comparison of variable configurations across scale lengths: An empirical study. Organizational and Psychological Measurement, 35, 319-324.
Number of Scale Points

Schwab, D. P., Heneman, H. H., & De Cotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 28, 549-562.

Scott, W. A. (1968). Attitude measurement. In G. Lindzey, & E. Aronson (Eds.), The handbook of social psychology. Reading, MA: Addison-Wesley.

Sedlacek, W. E. (1977). Test bias and the elimination of racism. Journal of College Student Personnel, 18(1), 16-20.
Ethnic Background

Segal, D. R., & Savell, J. M. (1975, August). Research on race relations in the U.S. Army: The multi-method matrix. In H. W. Sinciko, & L. A. Broedling (Eds.), Perspectives on attitude assessment: Surveys and their alternatives. Manpower Research and Advisory Services, Smithsonian Institution.
Ethnic Background

Shanker, P. (1977). Scaling trait-words by cross modality watching. Indian Psychological Review, 15(1), 1-6.

Shannon, R. H. (1981a). The validity of task analytic information to human performance research in unusual environments. In R. H. Shannon, & R. C. Carter (Eds.), Task analysis and the ability requirements of tasks: Collected Papers (NBDL-81R009). New Orleans, LA: Naval Biodynamics Laboratory.
Multiple-Choice Scales

Shannon, R. H. (1981b). Performance evaluation tests for environmental research (PETER) using task analysis. In R. H. Shannon, & R. C. Carter (Eds.), Task analysis and the ability requirements of tasks: Collected papers (NBDL-81R0009). New Orleans, LA: Naval Biodynamics Laboratory.
Rank Order Scales

Shannon, R. H. (1981c). The utility of task analytic techniques to research in unusual environments. In R. H. Shannon, & R. C. Carter (Eds.), Task analysis and the ability requirements of tasks: Collected papers (NBDL-81R009). New Orleans, LA: Naval Biodynamics Laboratory.
Rank Order Scales

Shannon, R. H. (1981d). A factor analytic approach to determining stability of human performance (NBDL-81R010). New Orleans, LA: U.S. Naval Biodynamics Laboratory.

Shannon, R. H. (1981e). Task analytic approach to human performance battery development. New Orleans, LA: U.S. Naval Biodynamics Laboratory.

Shannon, R. H., & Carter, R. C. (1981, September). Task analysis and the ability requirements of tasks: Collected papers (NBDL-81R009). New Orleans, LA: Naval Biodynamics Laboratory. (DTIC No. AD A111181)
Rank Order Scales

Sharon, A. T. (1980, May). An investigation of reference ratings of applicants for administrative law judge (PRR-80-6). Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development Center. (NTIS No. PB80-203573)

Sheehy, G. (1981). Pathfinders. New York: William Morrow and Company, Inc.
Length of Items and Number of Items

Shingledecker, C. A. (1983, June). Behavioral and subjective workload metrics for operational environments. AGARD-AMP Symposium: Sustained Intensive Air Operations: Physiological and Performance Aspects, Paris.
Future Research
Multiple-Choice Scales

Shosteck, H., & Fairweather, W. R. (1979). Physician response rates to mail and personal interviews surveys. Public Opinion Quarterly, 43, 206-217.
Interviewing

Silverstein, A. B. (1980, Summer). Item intercorrelations, item-test correlations, and test reliability. Educational and Psychological Measurement, 40(2), 353-355.

Sinaiko, H. W., & Broedling, L. A. (1975, August). Perspectives on attitude assessment: Surveys and their alternatives - Proceedings of a conference (TR-2). Washington, DC: Smithsonian Institution.

Singh, A. C., & Bilsbury, C. D. (1982). Scaling subjective variables by SPC (sequential pair comparisons). Behavioural Psychotherapy, 10(2), 128-145.

Sitton, L. R., Adams, I. G., & Anderson, H. N. (1980). Personality correlates of students' patterns of changing answers on multiple-choice tests. Psychological Reports, 47, 655-660.

Slater, P. E. (1962). Parent behavior and the personality of the child. Journal of Genetic Psychology, 101, 53-68.
Continuous and Circular Scales

Smith, K. F., & Baldauf, R. B., Jr. (1982). The concurrent validity of self-rating with interviewer rating in the Australian second language proficiency ratings scale. Educational and Psychological Measurement, 42, 1117-1124.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.
Behavioral Expectation Scales
Behaviorally Anchored Rating Scales

Smith, T. E. (1976, December). Scalable choice models. Journal of Mathematical Psychology, 14(3), 239-243.

Smith, T. W. (1981). Qualifications to generalize absolutes: "Approval of hitting" questions on the GSS. Public Opinion Quarterly, 45, 224-230.
Education
Gender
Wording of Items and Tone of Wording

Soeken, K. L., & Macready, G. B. (1982, September). Respondents' perceived protection when using randomized response. Psychological Bulletin, 92(2), 487-489.

Space, L. G. (1981). The computer as a psychometrician. Behavior Research Methods & Instrumentation, 13(4), 595-606.
Future Research

Spector, P. E. (1976). Choosing response categories for summated rating scales. Journal of Applied Psychology, 61(3), 374-375.

Spiers, P. A, & Pihl, R. O. (1976). The effect of study habits, personality, and order of presentation on success in an open-book objective examination. Teaching of Psychology, 3, 33-34.
Order of Items

Spies-Wood, E. (1980). Learned helplessness and item difficulty ordering. Psychologia Africana, 29-40.
Order of Items

Stager, P., & Paine, T. G. (1980). Separation discrimination in a simulated air traffic control display. Human Factors, 22(5), 631-636.

Staples, W. B., & Locander, W. B. (1975). Behaviorally anchored scales: A new tool for retail management evaluation and control. Journal of Retailing, 52(4), 39-95.
Behaviorally Anchored Rating Scales

Steinheiser, F. H., Jr., Epstein, K. I., Mirabella, A., & Macready, G. B. (1978, August). Criterion-referenced testing: A critical analysis of selected models (Technical Paper 306). College Park, MD: University of Maryland. (DTIC No. AD A061569)

Stewart, A. L., Ware, J. E., Jr., & Brook, R. H. (1977, March). A study of the reliability, validity, and precision of scales to measure chronic functional limitations due to poor health. Santa Monica, CA: The Rand Corporation. (DTIC No. AD A043261)

Stinson, J., & Stokes, J. (1980, June). How to multi-appraise. Management Today, 43-53.

Stouffer, S. A. (1955). Communism, conformity, and civil liberties. Garden City, NY: Doubleday.
Wording of Items and Tone of Wording

Strahan, R. F. (1980, October). More on averaging judges' ratings: Determining the most reliable composite. Journal of Consulting and Clinical Psychology, 48(5), 587-589.

Strahan, R. G. (1982). Assessing magnitude of effect from rank-order correlation coefficients. Educational and Psychological Measurement, 42(3), 763-765.

Strang, H. R. (1980). Effect of technically worded options on multiple-choice test performance. Journal of Educational Research, 73(5), 262-265.
Wording of Items and Tone of Wording

Straton, R. G., & Catts, R. M. (1980). A comparison of two, three and four-choice item tests given a fixed total number of choices. Educational and Psychological Measurement, 40, 357-365.

Subkoviak, M. J., & Roecks, A. L. (1976, Winter). A closer look at the accuracy of alternative data-collection methods for multidimensional scaling. Journal of Educational Measurement, 13(4), 309-317.

Swan, J. E., Epley, D. E. (1981, February). Completion and response rates for different forms of income questions in a mail survey. Perceptual and Motor Skills, 52(1), 219-222.

Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. Journal of Experimental Psychology, 7, 456-461.
Number of Scale Points

Sympson, J. B., Weiss, D. J., & Ree, M. J. (1982, August). Predictive validity of conventional and adaptive tests in an Air Force training environment (AFHRL-TR-81-40). Minneapolis, MN: Department of Psychology, University of Minnesota. (DTIC No. AD A119031)

Takane, Y. (1981, March). Multidimensional successive categories scaling: A maximum likelihood method. Psychometrika, 46(1), 9-28.

Tate, P. (1984). The blossoming of European AI. Datamation, 30(18), 85-88.
Future Research

Tatsuoka, K. (1979, July). Analytical test theory model for time and score (CERL Report E-8). Urbana, IL: Computer-Based Education Research Laboratory, University of Illinois. (DTIC No. AD A076601)

Tatsuoka, K. K., Birenbaum, M., Tatsuoka, M. M., & Baillie, R. (1981, August). Psychometric approach to error analysis on response patterns of achievement tests. Catalog of Selected Documents in Psychology, 11(58).

Tatsuoka, K. K., & Tatsuoka, M. M. (1982, Fall). Detection of aberrant response patterns and their effect on dimensionality. Journal of Educational Statistics, 7(3), 215-231.

Teel, K. S. (1978, July). Self-appraisal revisited. Personnel Journal, 364-367.

Thompson, B. (1980). Comparison of two strategies for collecting Q-Sort data. Psychological Reports, 47, 547-551.

Thompson, J. A., & Wilson, S. L. (1982, October). Automated psychological testing. International Journal of Man-Machine Studies, 17(3), 279-289.

Thorton, G. C. (1980). Psychometric properties of self-appraisals of job performance. Personnel Psychology, 33, 263-271.

Thurstone, L. L. (1927). A law of comparative judgment. Psychological Review, 34, 273-286.

Thurstone, L. L. (1929). Fechner's law and the method of equal-appearing intervals. Journal of Experimental Psychology, 12, 214-224.

Towsend, J. W. (1979). A Guttman analysis of scales developed by retranslation. Dissertation Abstracts International, 39(11-B), 5627-5628.

Trollip, S. R., & Anderson, R. I. (1982). An adaptive private pilot certification exam. Aviation, Space, and Environmental Medicine, 53(10), 992-995.
Future Research

Turney, J. R., & Cohen, S. L. (1976, June). The development of a work environment questionnaire for the identification of organizational problem areas in specific Army work settings (ARI Technical Paper-275). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (DTIC No. AD A028241)

Uhlaner, J. E., & Drucker, A. J. (1980). Military research on performance criteria: A change of emphasis. Human Factors, 22(2), 131-139.

Urry, V. W. (1977, August). Tailored testing: A spectacular success for latent trait theory (TS-77-2). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center. (NTIS No. PB 274 576)

Using behaviorally anchored rating scales (BARS) (1980). Small Business Report, 16-19.

Vale, C. D. (1981, August). Design annd implementation of a microcomputer-based adaptive testing system. Behavior Research Methods and Instrumentation, 13(4), 399-406.

Vance, R. J., Kuhnert, K. W., & Farr, J. L. (1978). Interview judgments: Using external criteria to compare bahavioral and graphic scale ratings. Organizational Behavior and Human Performance, 22, 279-294.
Interviewing

Van Heerden, J., & Hoogstraten, J. (1980, February). Response preference as a function of instructions in an unstructured questionnaire. Perceptual and Motor Skills, 50(1), 227-230.

van Rijn, P. (1980, June). Self-assessment for personnel examining: An overview (Personnel Research Report 80-14). Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development Center, Alternatives Task Force.
Ethnic Background

Verity, J. W. (1984). AI tools arrive in force. Datamation, 30(15), 44-53.
Future Research

Vidali, J. J. (1976). Reliability of the semantic differential under practical conditions. Psychological Reports, 39, 583-586.
Bipolar Scales
Semantic Differential Scales

Volans, P. J. (1982, October). Pros and cons of tailored testing: An examination of issues highlighted by experience with an automated testing system. International Journal of Man-Machine Studies, 17(3), 301-304.

Walizer, D. G., & Mietus, J. R. (1980, March). Development of an organizational survey feedback program for the 32D Air Defense Command (ARI-TR-433). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences. (AD A100 972)

Warm, T. A. (1978, December). A primer of item response theory (Technical Report 941278). Oklahoma City, OK: U.S. Coast Guard Institute. (DTIC No. AD A063072)
Future Research

Warmke, D. L., & Billings, R. S. (1979). A comparison of training methods for altering the psychometric properties of experimental and administrative performance ratios. Journal of Applied Psychology, 64, 124-131.

Weeks, M. F., & Moore, R. P. (1981). Ethnicity-of-interviewer effects on ethnic respondents. Public Opinion Quarterly, 45, 245-249.
Ethnic Background
Interviewing

Weiss, D. J. (1982, Fall). Improving measurement quality and efficiency with adaptive theory. Applied Psychological Measurement, 6(4), 473-492.

Welch, S., Comer, J., & Steinman, M. (1973). Interviewing in a Mexican-American community: An investigation of some potential sources of response bias. Public Opinion Quarterly, 37, 115-126.
Ethnic Background
Interviewing

Welsh, J. R. (1977, February). An investigation into the sources of halo error. Dissertation Abstracts International, 37(8-B), 4203.

Wherry, R. J., Sr., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. Personnel Psychology, 35, 521-551.

White-Blackburn, G. Blackburn, I. C., & Lutzker, J. R. (1980). The effects of objective versus subjective quiz items in a Psi course. Teaching Psychology, 7(3), 150-153.

Whitely, S. E. (1977). Relationships in analogy items: A semantic component of a psychometric task. Educational and Psychological Measurement, 37, 725-739.

Wiegand, D. (1983). Present standardization of test procedures for ergonomic test of tracked and wheeled vehicles of the Federal Armed Forces Germany. Federal Armed Forces Proving Ground 41. Unpublished paper.

Wienclaw, R. A., & Hines, F. E. (1982, November). A model for determining cost and training effectiveness tradeoffs. Training Equipment Interservice/Industry Training Equipmment Conference, 405-416.
Multiple-Choice Scales

Wilcox, R. R. (1982, Summer). Bounds on the k out of n reliability of a test, and an exact test for hierarchically related items. Applied Psychological Measurement, 6(3), 327-336.

Wilcox, R. R. (1982). Determining the length of multiple choice criterion-referenced tests when an answer-until-correct scoring procedures is used. Educational and Psychological Measurement, 42, 789-794.

Williams, J. A. (1964). Interviewer-respondent interaction: A study of bias in the information interview. Sociometry, 27, 338-352.
Interviewing

Wind, Y., & Lerner, D. (1979). On the measurement of purchase data: Surveys versus purchase diaries. Journal of Marketing Research, 16, 39-47.

Winkler, J. D., Kanouse, D. E., & Ware, J. E., Jr. (1982). Controlling for acquiescence response set in scale development. Journal of Applied Psychology, 67(5), 555-561.

Wise, S. L. (1982). A modified order-analysis procedure for determining unidimensional item sets. Dissertation Abstracts International, 42(7-A), 3121.

Wood, J. A. (1982). The quantification of verbal anchors used to denote occurrences of frequency, amount, and evaluation on five response category Likert scales. Dissertation Abstracts International, 43(2-A), 408-409.

Yadav, M. S., Govinda, R., & Thomas, K. T. (1976). Some psychometric studies in attitude scale construction. Psychological Studies, 21(1), 1-11.

Young, F. W., & Levinsoh, J. R. (1974). Two special-purpose programs that perform nonmetric multidimensional scaling. Behavior Research Methods and Instrumentation, 6(3), 354-355.

Zammuto, R. F., London, M., & Rowland, K. M. (1982). Organization and rater differences in performance appraisals. Personnel Psychology, 35, 643-658.

Zedeck, S. (1981). Behaviorally based performance appraisals. Aging and Work, 4(2), 89-100.

Zedeck, S., & Cascio, W. F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. Journal of Applied Psychology, 67(6), 752-758.

Zedeck, S., Jacobs, R., & Kafry, D. (1976). Behavioral expectations: Development of parallel forms and analysis of scale assumptions. Journal of Applied Psychology, 61, 112-115.

Zedeck, S., & Kafry, D. (1977, April). Capturing rater policies for processing evaluation data. Organizational Behavior and Human Performance, 18(2), 269-294.

Zedeck, S., Kafry, D., & Jacobs, R. (1976). Format and scoring variations in behavioral expectation evaluations. Organizational Behavior and Human Performance, 17, 171-184.

Multiple-Choice Scales
Questionnaire Layout

Appendix A

P-77-2

Questionnaire Construction Manual Annex Literature Survey and Bibliography

Table of Contents

Not every topic covered in P-77-2 is covered in this sequel. Appendix A provides the reader with a way to reference back to the original work of P-77-2, Questionnaire Construction Manual Annex. This may be useful for situations where researchers prefer to compare earlier questionnaire survey research with the more current literature.

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont.)

TABLE OF CONTENTS (Cont.)

Appendix B

Comparison Between P-77-2, Questionnaire Contruction Manual Annex, and the Sequel

This appendix delineates the content areas covered in P-77-2, and in this sequel. Each content area is identified by where it can be found in the sequel by title of the section, and then by what chapters it can be found in P-77-2. The usual case has been that a stand-alone section in the sequel can be found in more than one chapter of P-77-2. Some content areas are included in the sequel, but were not part of P-77-2. In addition, there are other content areas that were covered in P-77-2, but were not included in the sequel.

In P-77-2, common scaling techniques were found in two different chapters. Chapter III, Selection of Questionnaire Items to Be Used, compared various types of questionnaire items such as: ranking and rating scale items, paired comparisons, card sorts, semantic differential, checklists, multiple choice, and forced choice. Chapter IV compared the above mentioned scaling techniques. The sequel compares and updates the research for some of the same scaling techniques: multiple choice (this section includes Likert scales, Guttman scales, checklist, Q Sort, and behavioral scales), bipolar, semantic differential, rank order, and paired comparison.

A new addition to the literature has been included for the behavioral scales. The foundation for behavioral scales is the critical incident. The critical incident technique is mentioned in Chapter III of P-77-2. However, Behaviorally Anchored Rating Scales are not discussed in P-77-2. Behaviorally Anchored Rating Scales have been expanded so that there are now a wide variety of methods for this type of scale development and numerous forms of behavioral scales. The sequel includes sections on Behaviorally Anchored Rating Scales, Behavioral Expectation Scales, Behavioral Observation Scales, and Mixed Standard Scales. These behavioral scales were originally developed to encourage raters to observe behavior more accurately. They have been primarily used in questionnaire construction for performance appraisal purposes. Even so, there has been research which indicates that they have a broader application that includes surveys.

Format differences were discussed in P-77-2, Chapter III, Selection of Questionnaire Items, as to the pros and cons of various types of questionnaire items, in Chapter VIII, Considerations Related to the Physical Characteristics of Questionnaires, and in Chapter X, Characteristics of Respondents that Influence Questionnaire Results. In the sequel, questionnaire format has been addressed in Chapter VII as a stand-alone chapter.

A format difference for questionnaires has been included in the sequel in Section 7.2, Branching. Branching is a common approach used by researchers to guide respondents through a questionnaire to some questions, but not necessarily to all questions. Branching is also synonymous with other terms such as leading and routing. This topic area was not included in P-77-2.

P-77-2 incorporates the number of response alternatives to use in questionnaires, and the response anchoring, in Chapter VI, Number of Response Alternatives and Response Anchoring. The sequel separates out these two topic areas into three independent sections: Section 5.3, Number of Scale Points; Section 5.1, Response Alternatives; and Section 5.4, Middle Scale Point Position. The middle scale point position was incorporated into P-77-2 in Chapter VII, Order of Perceived Favorableness of Commonly Used Words and Phrases, and also in Chapter VI, Number of Response Alternatives and Response Anchoring.

The sequel includes a stand-alone section, Section 2.6, Continuous and Circular Scales, which was not part of P-77-2. Continuous scales have no scale points. The rationale for their use is that they will yield greater discrimination by raters. Circular scales are scales that were structured in a circumplex, or circle, to eliminate errors of extreme judgments, and errors of central tendency.

Open- and closed-end items are discussed in P-77-2 in Chapter II, Advantages and Disadvantages of Various Types of Questionnaires, and in Chapter III, Selection of Questionnaire Items to be Used. In the sequel, this topic area has been expanded and is found in Section 4.1, Open-Ended Items and Closed-End Items. Balancing response alternatives and the positive and negative wording of items is included in P-77-2 in Chapter V, Effects of Variation in Presentation of Questionnaire Items, in Chapter VI, Number of Response Alternatives and Response Anchoring, and in Chapter X, Characteristics of Respondents that Influence Questionnaire Results. These content areas are combined in the sequel to Chapter 4.5, Balanced Items.

The length of a questionnaire, the number of items in a questionnaire, and the number of words in an item are included in the sequel in Section 4.3, Length of Items and Number of Items. This content area is covered in P-77-2 in Chapter V, Effects of Variation in Presentation of Questionnaire Items, and in Chapter VIII, Considerations Related to the Physical Characteristics of Questionnaires. The ordering of items is found in P-77-2 in Chapter V, Effects of Variation in Presentation of Questionnaire Items, and in the sequel in Section 4.4, Order of Items. The sequel includes Section 4.2, Wording of Items and Tone of Wording. This content area may be found in P-77-2, Chapter V, Effects of Variation in Presentation of Questionnaire Items, and Chapter VII, Order of Perceived Favorableness of Commonly Used Words and Phrases.

Interviewing is treated as an independent topic in the sequel in Section 6.1. In P-77-2, interviewing is discussed in Chapter II, Advantages and Disadvantages of Various Types of Questionnaires, and in Chapter IX, Considerations Related to the Administration of Questionnaires.

In recent years, there has been a trend away from research that focuses on questionnaire construction for content areas such as: format, number of scale points, and types of response alternatives. A greater focus has been placed on respondent demographic characteristics that might influence a rating, training respondents in how to rate, the complexity of the questionnaire, and the cognitive complexity of the rater. The theory of cognitive complexity has its foundation in the work of Kelly (1955), and has been defined as the ability to differentiate person-objects in the

social environment. The sequel includes Section 6.2, Cognitive Complexity. This topic is not covered in P-77-2.

Demographic characteristics that describe respondents in questionnaire construction have been divided into four sections in the sequel, and these sections are: Section 6.3, Education; Section 6.4, Ethnic Background; Section 6.5, Gender; and Section 6.6, Age. These demographic characteristics are found in P-77-2 in Chapter IX, Considerations Related to the Administration of Questionnaires, and In Chapter X, Characteristics of Respondents that Influence Questionnaire Results.

P-77-2 and the sequel both cover the same content areas in most instances. There are some areas where there is not overlap. For example, P-77-2 does not include material on Behaviorally Anchored Rating Scales, Continuous and Circular Scales, and Cognitive Complexity, while the sequel does. The sequel does not include areas on Questionnaire Administration and Evaluation of Questionnaire Results, which are included in P-77-2. (See Appendix C for an overview of content areas covered by P-77-2 and the sequel.)

Appendix C

Overview of Content Areas Covered by P-77-2 and the Sequel

Questionnaire Construction Content Areas	P-77-2	Sequel
Scaling Techniques	Yes	Yes
Behaviorally Anchored Rating Scales	No	Yes
Format	Yes	Yes
Branching	No	Yes
Response Alternatives	Yes	Yes
Continuous and Circular Scales	No	Yes
Open- and Closed-End Items	Yes	Yes
Balancing Response Alternatives	Yes	Yes
Wording of Items	Yes	Yes
Length and Number of Items	Yes	Yes
Order of Items	Yes	Yes
Interviewing	Yes	Yes
Cognitive Complexity	No	Yes
Demographic Characteristics	Yes	Yes
Administration of Questionnaires	Yes	No
Evaluation of Questionnaire Results	Yes	No
Further Research	Yes	Yes

Appendix D

Future Research Recommendations

Future research recommendations have been grouped together according to content area by chapter: Chapter II, Scale Categories; Chapter III, Behavioral Scales; Chapter IV, Design of Questionnaire Items; Chapter V, Design of Scale Categories; Chapter VI, Interviewer and Respondent Characteristics; and Chapter VII, Questionnaire Format. These groupings are not ranked. The intention is to provide an overview of each content area for what types of research might be performed. These recommendations were selected because of perceived gaps in survey research for specific content areas.

Chapter II. Scale Categories

- Practical, workable procedures are sorely needed to focus investigators on refining and implementing developmental procedures in the construction of questionnaires, instead of arbitrarily borrowing items from other surveys. Development and validation of such procedures should be addressed.

- No one scale category can be recommended over another. More research is required to replicate studies in order to establish which is the best scale category to use for specific types of applications.

- The combination of scales, such as ordinal and interval scales, used in conjoint measurement requires further refinement.

- Guttman scaling theory requires further exploration as it applies to developing interviews which are predictive of future behavior, physical health, etc. These scales are probably the most difficult to develop of all the scaling techniques covered in this report.

- Research is needed to find ways to develop Guttman scales using fewer subjects and in shorter time frames. This would be a major breakthrough for expanding the use of Guttman scales.

- More studies are needed to determine whether subjects are confounding trait dimensions with response alternatives.

- Replication of studies is required to determine whether subjects are influenced by descriptive anchors, thereby making greater use of the categories at the extreme ends of the scale.

- Further evidence is required to substantiate the existence of response style as it relates to the order of item presentation.

- It cannot be concluded that adverb and numerical scales measure meaning in the same way. Factor analysis would allow for a better understanding of the meaning of the items. Additional studies to confirm/disconfirm the findings would be of value.

- More research will be required to determine reliable and valid procedures for transforming ordinal data into interval scales in conjoint measurement.

- Data collection for value profiles has more or less abandoned the rank order technique for the more popular Likert scaling method. Yet, there is evidence to support the use of rank ordering or paired comparison methods as being more reliable. Test-retest reliability for the scales mentioned above needs further investigation by comparing rho and Kendall's tau in the statistical analysis. In addition, software could be developed which would make this technique easier and more quickly performed.

- Future studies which compare rank order methods with other scale methods should consider experimental designs with repeated measures since an individual's preference for type of scale may have affected the stability of the results over time. Experimental designs should focus on the effects of individuals, the effects of method, and the individual method interaction effects.

- Researchers have extended the use of card sorts for marketing purposes. It was found that subjects were able to reconstruct the underlying situational hierarchical structures on a card sort for the first trial. Second trial card sorts shifted to a less ordered structure. Further investigation will be required to resolve why this type of shift was observed.

- Continuous scales have been used to provide greater discrimination by respondents. However, more research is needed since there has been evidence that respondents may be rating continuous scales on the equivalent of five or six categories depending on what is being measured.

- Continuous scales have been integrated and transformed into category scales in the measurement of psychophysical phenomena. The combination of these two scales requires more research for measuring vehicle environments, such as vibration or temperature.

Chapter III. Behavioral Scales

- The application of Behaviorally Anchored Rating Scales (BARS) to measure group morale was encouraging as an alternative to the traditional self-report measures obtained in surveys. Other applications for BARS surveys may be useful in reducing the subjectivity found in other self report instruments.

- When BARS have been used solely for performance appraisal, the time and cost factors involved in their development have been high. Research is needed to identify multiple uses for BARS, such as delineating organizational goals, in order to make development efforts worthwhile.

- Further research is needed to validate rater training programs using BARS to increase rater accuracy.

- Further research is needed to investigate the possible increase in validity for BARS when multiple rater groups stipulate the dimensions of performance that effects them.

- Most empirical research dealing with performance evaluation systems, such as Behavioral Expectation Scales (BES), measures improvements in attitudes and performance for short study periods of less than one year. Longitudinal studies for the impact of BES over time would be worthwhile in establishing or disputing the long-range results that these systems hope to have.

- BES require further investigation as to behavior change, relationship to performance effectiveness, and developmental and scoring procedures.

- More studies are needed to test the effects of various training interventions on psychometric error, and the use of BES.

- Behavioral Observation Scales require more research to determine whether these scales are really measuring simple observations or whether they are measuring trait-like judgment due to rater recall over time.

- Mixed Standard Scales (MSS) research should focus on the rating process and environment since the psychometric properties of the rating scales may be influenced by organizational reward practices and organizational climate.

- MSS have been shown to be reliable and valid. However, raters prefer other scale types over MSS. Additional research is needed to determine whether MSS are appropriate for use in operational test and evaluation (OT&E).

- Anchor items on MSS require investigation for possible multidimensionality since this would yield inconsistent ratings.

Chapter IV. Design of Questionnaire Items

- Further investigation is needed to substantiate whether respondents with more education will answer open-ended and closed-end forms consistently. Respondents with less education do not respond consistently to these form differences.

- There has not been consistent replication as to the wording of items and the tone of wording so that researchers have not been able to predict when the items will be influenced by the wording and which words will influence the results. More research is needed in this area.

- Researchers need to assess various procedures or methods as a way to identify the use of specific words which are used in items. Such a method may have the potential to identify the structure of the item itself.

- Research is needed in the OT&E community to determine whether item reduction techniques used in marketing surveys and behavioral scale development would be applicable in reducing the number of items used in field surveys.

- Methodological research for understanding why order effects occur is an important area for questionnaire construction research to examine. For example, it is assumed that separating two items by several or many others may eliminate a known context effect. Further investigation and greater understanding of these issues is needed.

- More needs to be known about the usefulness of balancing anchors in conjunction with what type of application the scale will have.

- Replication would be helpful to indicate whether items of greater length will elicit responses toward the middle of the scales. In addition, replication is required to determine whether items of short length elicit responses toward the positive end of the scale.

- Further replication is suggested for long items which are negatively worded to determine whether they elicit responses toward the midrange of the scale.

- More studies would be desirable to investigate both positive and negative items on a questionnaire. There has been some evidence that negatively worded items may result in less accurate responses, and reduce response validity.

Chapter V. Design of Scale Categories

- There has been some evidence that formats which are easiest to rate would be best for respondents who have lower levels of education. Further research is needed to explore this finding, and to identify which formats those might be.

- If response alternatives are selected independent of the item, and measured for bands along the scale dimension, there may be the possibility that response alternative linkage to the item may modify the standard deviation of each response. The linkage of the response alternative to the item for variations in the standard deviation needs further investigation.

- More studies would be useful in examining whether fully labeled scales yield less skewed responses than scales anchored only at the extremes. There has been some evidence that fully labeled scales achieved higher test-retest correlations. This finding was observed with market-segmentation studies, and would need investigation for generalizing to OT&E studies.

- There has been contradictory evidence that verbal anchors may lower accuracy in psychophysical task ratings. More thorough and systematic research would be useful in establishing the effects of verbal anchoring on psychometric criteria for psychophysical tasks.

- Pictorial anchors have been subjected to limited investigation. This methodology needs to be extended to different types of visually perceived stimuli in OT&E.

- There has been conflicting evidence regarding individuals who may or may not respond to the "Don't Know" category. One theory has proposed that there is a unique constellation of traits which describes individuals who "float" back and forth between the "Don't Know" response, and other response alternatives. "Floating" may not have a single cause, although there is no model to describe subjects who exhibit this behavior. Additional investigation is needed.

- Further studies are needed to identify the respondent characteristics and/or the processes involved in responses to the presence or absence of a middle alternative.

Chapter VI. Interviewer and Respondent Characteristics

- More studies would be useful to improve interview techniques which are used in telephone surveys. Experimental variations of question presentation may be useful.

- Military survey research that incorporates interviews needs to investigate ways to obtain more lead time in survey development, increase response rate for large periodic surveys, control standardization in field administration, and control for response bias of personnel brought about by the influence of superiors.

- More research is needed on the effects of third parties on face-to-face interviews.

- To investigate cognitive complexity, the contextual differences for the type of organization, and the characteristics of the respondents, might account for the failure to replicate previous results. Organizations other than institutions of high education may be a more appropriate environment to investigate whether cognitive complexity is a relevant variable in the rating task. Cognitive complexity issues should be investigated for OT&E respondent populations.

- Further replication is needed on the generalizability of item rating by level of education from the general U.S. population to military personnel.

- There may be an interaction among educational level of respondents, response consistency in rating items, and intensity of feeling about items. Investigation of this interaction is needed to replicate previous research.

- Educational level and item nonresponse may be related; this area could use more research. This type of survey is difficult to design since it is challenging to obtain strategies for identifying nonrespondents.

- Very little is known about the effect of education on questionnaires that are designed for use in performance appraisal. Research in this area may be useful.

- There has not been consistency in findings for the impact of education and questionnaire construction. More research is needed on the influence of combined relationships with other demographic variables, as well as generalizability of these results to military personnel.

- Results have been mixed as to ethnic background of raters assessing ratees on performance appraisal scales. Further research is needed for the effect of same-race raters and different-race raters.

- Further research may be required to determine the extent to which self assessment can be used, and the influence of ethnic background on self assessment.

- Not enough is known regarding the impact of ethnic background on questionnaire construction. More studies for the influence of ethnic background and questionnaire construction would be useful in designing and analyzing these instruments.

- Review of the research relating to questionnaire construction and the differentiation of response patterns by gender of respondents has received mixed results. Some studies have found differences in rating by females and males, while other studies have not. Research is needed to identify interactions of gender, education, age, and ethnic background. For example, investigation of item content and gender may be useful since the content of an item may have the potential to bias survey results where males and females hold different values.

- For opinion surveys, further studies would be useful in examining the influence of age on item content to assess the historical perspective of the different groups.

- Nonresponses to items following a branch, and survey nonresponse, may be influenced by age and education of the respondent. Replication would be beneficial in assessing this phenomenon as it is influenced by demographic characteristics of Army populations in OT&E.

Chapter VII. Questionnaire Format

- Evidence supporting any one format was sparse and inconsistent. More studies may be useful in identifying strengths and weaknesses of different scale formats. However, other methodological issues may have greater potential for research such as: Rigor of developmental procedures, preference of format by raters, matching characteristics of rater and format, assessment of scoring systems, and examination of different sampling techniques.

- A major implication for future research appears to be the reassessment of scale item selection. This could be approached in two ways. First, through a technique that would allow for the wording of each item (e.g., the Echo method). Second, item reduction procedures that maintain contruct validity.