

## Questionnaire Construction Manual

This Section contains copies of the Army's Questionnaire Construction Manual, an example of using scaled questionnaire analysis techniques, a Navy Human Factors in Operational System Testing Manual, and a Naval Postgraduate School paper on the use of interval scales using data from categorical judgments. These documents provide the JT&E analysts an means to convert normally qualitative human source data into quantitative data by using scaled questionnaire techniques. The analyst using these reference should **NOT** use the material in a quick look fashion. The skill related to developing a questionnaire that is not biased or leading can only be developed through a thorough understanding of the concepts and processes related to developing scaled questionnaires. As a simple example, a question that requires a "yes" or "no" response is automatically leading and could be biased. This is because the question must be stated in either a positive or negative sense. If the developer is interested in the positive aspects of a TTP or process, they often will write a question a manner that psychologically requires the respondent to respond in a positive manner. Scaled questionnaires result in a spread of positive and negative. The Questionnaire Construction Manual also provides a number of questionnaire response alternatives and their associated standard deviations from the norm. Careful selection of alternative responses can yield a highly informative data collection tool that be evaluated with a number of non-parametric tests. If properly constructed, the JT&E analyst can construct questionnaires that contain two or more categories or conditions for which the respondents are providing data. These categories can be evaluated independently or as a group (after the application of non-parametric statistics to determine if a group of question results has the same underlying statistical population. If so, it may be possible to combine the questionnaire groups.

The Questionnaire Construction manual also contains procedures and processes that should be used when administrating questionnaires and surveys to potential respondents. The remaining manuals also provides information about questionnaire construction and application. It is noted that an interval scale, e.g. 1 to 10, is not as good as a scaled response question e.g six responses available. This is because it is difficult for people to reliably select a specific number that may correspond to the requested question. It is also prudent to "test" the questionnaire with a group of nonparticipants to identify any deficiencies in the questions asked in the questionnaire. This provides an opportunity to resolve difficulties without causing confusion at the time of administering the questionnaire to the intended population. Great care must be taken to ensure that the questionnaire process is not contaminated.

Regarding the identification of criteria related to questionnaires. If scaled questionnaires are used, the analyst can derive a criterion that is based on the number of alternative selections. For example, six selections such as: Completely Agree - 6.0 score, Strongly Agree- 5.0 score, Generally Agree - 4.0 score, Generally Disagree - 3.0 score, Strongly Disagree - 2.0 score, and Completely Disagree - 1.0 score has a mean midpoint value of 3.5 (value midpoint of  $(4 + 3)/2$ ). This midpoint value can represent a minimally acceptable criteria e.g. Less than 3.5, represents disagree. 3.5 or greater represents agreement. The analyst can also expand the analysis to each of the categories above or below 3.5 (e.g. how many scored greater than 5.5?). See the example for more detail.





**Research Product 89-20**

# **Questionnaire Construction Manual**

**June 1989**

**Fort Hood Field Unit  
Systems Research Laboratory**

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

Approved for public release; distribution is unlimited.



# U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction  
of the Deputy Chief of Staff for Personnel

**EDGAR M. JOHNSON**  
Technical Director

**JON W. BLADES**  
COL, IN  
Commanding

---

Research accomplished under contract  
for the Department of the Army

Essex Corporation

Technical review by

David Meister  
W.F. Moroney, MSC, U.S. Navy

## NOTICES

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

**NOTE:** This Research Product is not to be construed as an official Department of the Army  
document, unless so designated by other authorized documents.



UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

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) --		5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Research Product 89-20			
6a. NAME OF PERFORMING ORGANIZATION Essex Corporation Human Factors & Training Systems Group		6b. OFFICE SYMBOL (if applicable) --		7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences	
6c. ADDRESS (City, State, and ZIP Code) 741 Lakefield Road, Suite B Westlake Village, CA 91361		7b. ADDRESS (City, State, and ZIP Code) ARI Field Unit at Fort Hood 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) PERI-2A		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. 0
11. TITLE (Include Security Classification) Questionnaire Construction Manual					
12. PERSONAL AUTHOR(S) Babbitt, Bettina A. (Essex Corporation), and Nystrom, Charles O. (ARI)					
13a. TYPE OF REPORT Final		13b. TIME COVERED FROM 84/12 TO 85/03		14. DATE OF REPORT (Year, Month, Day) 1989, June	
				15. PAGE COUNT 227	
16. SUPPLEMENTARY NOTATION (Continued) This is a revised version of the July 1976 Questionnaire Construction Manual, P-77-1, originally authored by R. F. Dyer, J. J. Mathews, C. E. Wright, and K. L. Yudowitch.					
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)		
FIELD	GROUP	SUB-GROUP	Questionnaire construction      Scaling techniques		
			Questionnaire administration      Response anchoring		
			Attitude scales (Continued)		
19. ABSTRACT (Continue on reverse if necessary and identify by block number)  This Questionnaire Construction Manual is a revised version of a 1976 manual. The latest research methods for developing questionnaires are presented. The manual was designed to guide individuals who develop and/or administer questionnaires as part of Army field tests and evaluations. The content is applicable to many nonmilitary applications.					
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT <input type="checkbox"/> UNCLASSIFIED/UNLIMITED <input checked="" type="checkbox"/> SAME AS RPT <input type="checkbox"/> DTIC USERS			21. ABSTRACT SECURITY CLASSIFICATION Unclassified		
22a. NAME OF RESPONSIBLE INDIVIDUAL Charles O. Nystrom			22b. TELEPHONE (Include Area Code) (817) 288-9118		22c. OFFICE SYMBOL PERI-SH

DD Form 1473, JUN 86

Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED

ARI Research Product 89-20

16. SUPPLEMENTARY NOTATION (Continued)

Operations Research Associates, Palo Alto, CA. Charles O. Nystrom is the Contracting Officer's Representative.

18. SUBJECT TERMS (Continued)

Response alternatives  
Pretesting questionnaires  
survey interviews

**Research Product 89-20**

# **Questionnaire Construction Manual**

**Bettina A. Babbitt**

Essex Corporation

and

**Charles O. Nystrom**

U.S. Army Research Institute

**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 Evaluation**

Approved for public release; distribution is unlimited.



## FOREWORD

---

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI), Field Unit at Fort Hood, Texas, actively guided this revision of their 10-year-old Questionnaire Construction Manual (P-77-1). The questionnaire construction manual was designed to guide individuals who develop and/or administer questionnaires as part of Army operational tests. It is, however, suitable for a variety of disciplines and occupations. Guidance is provided in the development of questionnaire items, administration procedures, types of questionnaire items, attitude scales and scaling techniques, response anchoring and response alternatives, format considerations, pretests, interviews, demographic characteristics, and evaluation of results.

This product was completed under Program Task 1.5.1, "Soldier/System Considerations in Force Development User Testing (Advanced Development)." ARI and the Sponsor for the product work under a "Memorandum of Agreement between ARI and Training and Doctrine Command (TRADOC) Combined Arms Test Activity (TCATA)" that was signed in May 1981. The Chief of TCATA's Methodology and Analysis Section has been briefed on the product content. TCATA has been using the predecessor Questionnaire Construction Manual to test officers for over 10 years and would like to use the updated product.



EDGAR M. JOHNSON  
Technical Director

## ACKNOWLEDGMENTS

---

Several people helped to prepare this manual. A 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 contribution of Mr. George M. Gividen, U.S. Army Research Institute for the Behavioral and Social Sciences, is most gratefully acknowledged. Mr. Clarence A. Semple, Essex Corporation, contributed generously in editing. Mrs. Joan M. Funk, Essex Corporation, merits special recognition for her technical assistance in preparing and editing the manuscript.

# QUESTIONNAIRE CONSTRUCTION MANUAL

## EXECUTIVE SUMMARY

---

This manual updates the 10-year-old "Questionnaire Construction Manual." The revision was prepared primarily by the Essex Corporation under contract to the Army Research Institute for the Behavioral and Social Sciences (ARI). It has the same purpose as the earlier version--to provide guidance for those who construct and/or administer questionnaires as part of Army operational tests and evaluations such as those conducted by the TRADOC Combined arms Test Activity and the Operational Test and Evaluation Agency. Much of the content is applicable to more than operational test situations; the manual should prove useful to all persons involved in the construction and administration of surveys, interviews, or questionnaires.

In 1975, Operations Research Associates reviewed the research literature on the construction and administration of questionnaires and interviews. They produced two products. One was the forerunner of this manual. It was titled "Questionnaire Construction Manual" and was published by ARI in 1976. A revision was done in 1976 and issued in quantity in 1977 as ARI Special Publication P-77-1. The other product was a report of the literature survey and a bibliography of the articles examined. It was issued in 1977 as P-77-2, with the title "Questionnaire Construction Manual Annex: Literature Survey and Bibliography."

In 1983, the literature was again reviewed, but only from the point where ORA's review had ended in 1975. Analysis of the more recent literature provided the basis for the revision to the manual. A report of the literature survey has been published under the title, "Questionnaires: Literature Survey and Bibliography."



# QUESTIONNAIRE CONSTRUCTION MANUAL

## CONTENTS

---

	Page
I. INTRODUCTION . . . . .	1
A. Purpose and Organization of This Manual . . . . .	1
B. Definition of Questionnaire . . . . .	2
C. Conventions Used in This Manual . . . . .	3
D. Keeping This Manual Up to Date . . . . .	4
E. Reporting Problems and Suggestions for Improvement . . . . .	5
II. MAJOR QUESTIONNAIRE TYPES AND ADMINISTRATION PROCEDURES . . . . .	7
A. Overview . . . . .	7
B. Types of Questionnaires Discussed in This Manual . . . . .	8
C. Ways That Questionnaires Can Be Administered . . . . .	9
D. Structured Interviews Versus Other Types of Questionnaires . . . . .	11
III. CONTENT OF QUESTIONNAIRE ITEMS . . . . .	13
A. Overview . . . . .	13
B. Determining Questionnaire Content Preliminary Research . . . . .	14
C. Other Considerations Related to Questionnaire Content . . . . .	20
IV. TYPES OF QUESTIONNAIRE ITEMS . . . . .	23
A. Overview . . . . .	23
B. Open-Ended Items . . . . .	24
C. Multiple Choice Items . . . . .	28
D. Rating Scale Items . . . . .	32
E. Behavioral Scale Items . . . . .	37
F. Ranking Items . . . . .	44
G. Forced Choice Items . . . . .	47
H. Card Sorting Items/Tasks . . . . .	50
I. Semantic Differential Items . . . . .	52
J. Other Types of Items . . . . .	55
V. ATTITUDE SCALES AND SCALING TECHNIQUES . . . . .	59
A. Overview . . . . .	59
B. Thurstone Scales . . . . .	61
C. Likert Scales . . . . .	64
D. Guttman Scales . . . . .	68
E. Other Scaling Techniques . . . . .	71

CONTENTS (Continued)

	Page
VI. PREPARATION OF QUESTIONNAIRE ITEMS . . . . .	73
A. Overview . . . . .	73
B. Mode of Items . . . . .	74
C. Wording of Items . . . . .	75
D. Difficulty of Items . . . . .	91
E. Length of Question/Stem . . . . .	94
F. Order of Question/Stems . . . . .	95
G. Number of Response Alternatives . . . . .	98
H. Order of Response Alternatives . . . . .	102
VII. RESPONSE ANCHORING . . . . .	105
A. Overview . . . . .	105
B. Types of Response Anchors . . . . .	106
C. Anchored Versus Unanchored Scales . . . . .	109
D. Amount of Verbal Anchoring . . . . .	110
E. Procedures for the Selection of Verbal Scale Anchors . . . . .	111
F. Scale Balance, Midpoints, and Polarity . . . . .	112
VIII. EMPIRICAL BASES FOR SELECTING MODIFIERS FOR RESPONSE ALTERNATIVES . . . . .	115
A. Overview . . . . .	115
B. General Considerations in the Selection of Response Alternatives . . . . .	118
C. Selection of Response Alternatives Denoting Degrees of Frequency . . . . .	132
D. Selection of Response Alternatives Using Order of Merit Lists of Descriptor Terms . . . . .	133
E. Selection of Response Alternatives Using Scales Values and Standard Deviations . . . . .	135
F. Sample Sets of Response Alternatives . . . . .	156
IX. PHYSICAL CHARACTERISTICS OF QUESTIONNAIRES . . . . .	163
A. Overview . . . . .	163
B. Location of Response Alternatives Relative to the Stem . . . . .	164
C. Questionnaire Length . . . . .	166
D. Questionnaire Format Considerations . . . . .	168
E. Use of Answer Sheets . . . . .	172
F. Use of Branching . . . . .	173
X. CONSIDERATIONS RELATED TO QUESTIONNAIRE ADMINISTRATION . . . . .	175
A. Overview . . . . .	175
B. Instructions . . . . .	176
C. Anonymity for Respondents . . . . .	178

CONTENTS (Continued)

	Page
D. Motivational Factors . . . . .	183
E. Administration Time . . . . .	186
F. Characteristics of Administrators . . . . .	187
G. Administration Conditions . . . . .	188
H. Training of Field Test Evaluators . . . . .	189
I. Other Factors Related to Questionnaire Administration . . . . .	191
 XI. PRETESTING OF QUESTIONNAIRES . . . . .	 193
A. Overview . . . . .	193
B. Guidelines for Pretesting Questionnaires . . . . .	194
 XII. CHARACTERISTICS OF RESPONDENTS THAT INFLUENCE QUESTIONNAIRE RESULTS . . . . .	 197
A. Overview . . . . .	197
B. Social Desirability and Acquiescence Response Sets . . . . .	198
C. Other Response Sets or Errors . . . . .	200
D. Effects of General Attitudes of Respondents . . . . .	203
E. Effects of Demographic Characteristics on Responses . . . . .	204
 XIII. EVALUATING QUESTIONNAIRE RESULTS . . . . .	 207
A. Overview . . . . .	207
B. Scoring Questionnaire Responses . . . . .	208
C. Data Analyses . . . . .	210
 XIV. INTERVIEW CONSIDERATIONS . . . . .	 211
A. Overview . . . . .	211
B. Structured and Unstructured Interviews . . . . .	212
C. Interviewer's Characteristics Relative to Interviewee . . . . .	213
D. Situational Factors . . . . .	215
E. Training Interviewers . . . . .	217
F. Data Recording and Reduction . . . . .	218
G. Special Interviewer Problems . . . . .	219

LIST OF TABLES

Table VIII-B-1. Words considered unratable by subjects . . . . .	119
VIII-B-2. Words evoking bimodality of response . . . . .	120
VIII-B-3. Sample list of phrases denoting degrees of acceptability . . . . .	122

	Page
Table VIII-B-4. A second sample list of phrases denoting degrees of acceptability . . . . .	122
VIII-B-5. Candidate midpoint terms' scale values and standard deviations as determined by several different studies . . . . .	124
VIII-C-1. Degrees of frequency . . . . .	132
VIII-D-1. Order of merit of selected descriptive terms . . . . .	133
VIII-D-2. Order of merit of descriptive terms using "use" as a descriptor . . . . .	134
VIII-E-1. Acceptability phrases . . . . .	136
VIII-E-2. Degrees of excellence: First set . . . . .	137
VIII-E-3. Degrees of excellence: Second set . . . . .	138
VIII-E-4. Degrees of like and dislike . . . . .	139
VIII-E-5. Degrees of good and poor . . . . .	140
VIII-E-6. Degrees of good and bad . . . . .	141
VIII-E-7. Degrees of agree and disagree . . . . .	142
VIII-E-8. Degrees of more and less . . . . .	143
VIII-E-9. Degrees of adequate and inadequate . . . . .	144
VIII-E-10. Degrees of acceptable and unacceptable . . . . .	145
VIII-E-11. Comparison phrases . . . . .	147
VIII-E-12. Degrees of satisfactory and unsatisfactory . . . . .	148
VIII-E-13. Degrees of unsatisfactory . . . . .	148
VIII-E-14. Degrees of pleasant . . . . .	149
VIII-E-15. Degrees of agreeable . . . . .	149
VIII-E-16. Degrees of desirable . . . . .	150
VIII-E-17. Degrees of nice . . . . .	150
VIII-E-18. Degrees of adequate . . . . .	151

CONTENTS (Continued)

---

	Page
Table VIII-E-19. Degrees of ordinary . . . . .	151
VIII-E-20. Degrees of average . . . . .	152
VIII-E-21. Degrees of hesitation . . . . .	152
VIII-E-22. Degrees of inferior . . . . .	153
VIII-E-23. Degrees of poor . . . . .	153
VIII-E-24. Descriptive phrases . . . . .	154
VIII-F-1. Sets of response alternatives selected so phrases are at least one standard deviation apart and have parallel wording . . . . .	157
VIII-F-2. Sets of response alternatives selected so that intervals between phrases are as nearly equal as possible . . . . .	159
VIII-F-3. Sets of response alternatives selected from lists giving scale values only . . . . .	161
VIII-F-4. Sets of response alternatives selected using order of merit lists of descriptor terms . . . . .	162

LIST OF FIGURES

Figure IV-B-1. Examples of open-ended items . . . . .	24
IV-C-1. Examples of multiple choice items . . . . .	29
IV-D-1. Examples of numerical rating scale items . . . . .	32
IV-D-2. Example of graphic rating scale item . . . . .	33
IV-D-3. Examples of discrete and continuous scales used to rate perception of tones . . . . .	34
IV-E-1. Example of BARS's seven dimensions describing technician behavior . . . . .	39
IV-E-2. Example of BARS items representing performance and effort on the job . . . . .	40
IV-E-3. Example of BOS item representing description of foreman's job . . . . .	41

CONTENTS (Continued)

---

	Page
Figure IV-E-4. Example of MSS items representing highway patrol stopping vehicles for violations . . . . .	41
IV-F-1. Examples of ranking items . . . . .	45
IV-G-1. Examples of forced choice items . . . . .	48
IV-I-1. Examples of semantic differential items . . . . .	53
IV-J-1. Examples of checklists . . . . .	55
IV-J-2. Example of checklist pertaining to equipment problems . .	56
IV-J-3. Examples of formats providing for supplementary responses . . . . .	58
VI-C-1. Example of question form and incomplete statement form of stem . . . . .	76
VI-C-2. An insufficiently detailed question stem, plus revision . . . . .	78
VI-C-3. Examples of loaded questions . . . . .	81
VI-C-4. Examples of leading questions . . . . .	82
VI-C-5. Example of a threatening question . . . . .	83
VI-C-6. Example of a question asking the respondent to criticize . . . . .	84
VI-C-7. Examples of compound questions and alternatives . . . . .	85
VI-C-8. Example of ambiguous question and alternative . . . . .	86
VI-C-9. Example of ambiguity of wording . . . . .	87
VI-C-10. Alternate ways of expressing directionality and intensity . . . . .	89
VI-D-1. Example of hard to understand item and alternative . . .	91
VI-F-1. Example of Bradley Fighting Vehicle Questionnaire for multiple groups . . . . .	96
VI-H-1. Example of rating scale item with alternate ordering of response alternatives . . . . .	104

CONTENTS (Continued)

---

	Page
Figure VII-B-1. Types of response anchors . . . . .	107
VII-F-1. Examples of scale balance, midpoints, and polarity . . .	113
VIII-B-1. Inclusion of the "Don't Know" response alternative for a maintenance vehicle questionnaire . . . . .	128
VIII-B-2. Two formats using "outstanding" and "superior" . . . .	130
VIII-B-3. Response alternatives frequently recommended by ARI . .	131
IX-B-1. Arrangement of items with same rating scale response alternatives . . . . .	165
IX-D-1. Original questionnaire format and modified questionnaire format . . . . .	170
X-C-1. An example of a Privacy Act statement . . . . .	182



Chapter I: IntroductionA. Purpose and Organization of This Manual1. Purpose

This manual has been prepared primarily for the use and guidance of those who are tasked to develop and/or administer questionnaires as part of Army field tests and evaluations, such as those conducted by the TRADOC Combined Arms Test Activity (TCATA), the Combat Developments Experimentation Command (CDEC), the Operational Test and Evaluation Agency (OTEA), and the several Army Boards and Schools. The general content and concepts, however, are applicable to a variety of situations. As such, the manual should prove useful to all individuals involved in the construction and administration of surveys, interviews or questionnaires.

2. Organization

Information and guidance relating to the preparation of items for questionnaires and for their assembly and arrangement into a complete questionnaire are presented in Chapters II through X. Chapter XI discusses the importance of, and procedures for, pretesting questionnaires prior to their regular administration. Chapter XII discusses characteristics of respondents that influence questionnaire results. The analysis and evaluation of responses to a questionnaire are briefly dealt with in Chapter XIII. Finally, a number of considerations regarding the presentation of questions by means of an interview are discussed in Chapter XIV.

B. Definition of Questionnaire

As used in this manual, the word "questionnaire" refers to an ordered arrangement of items (questions, in effect) intended to elicit the evaluations, judgments, comparisons, attitudes, beliefs, or opinions of personnel. The content and format of the items may vary widely. A visual mode of presenting the items is employed. In the past, this meant that the items were typed or printed on paper, but now items can also be presented by closed circuit television or on a cathode ray tube (CRT) or on a video display terminal (VDT) under the control of a computer program. If the items are first read by an interviewer and then given verbally to the respondent, the questionnaire may also be termed a "structured interview." Hence, questionnaires and interviews have some common properties. Questionnaire items used to be responded to by scribing words or marks with a pen or pencil, but this aspect too has been enlarged to include typed, punched, button-pushing, light-penned, joystick, and verbal responses.

While questionnaires are "data collection forms," not all data collection forms are questionnaires. Those forms used by personnel to enter instrument readings or to record their counts or observations (e.g., time of first detection, number of targets correctly identified, number of rounds fired) are not directly addressed in this manual.

C. Conventions Used in This Manual

1. Identification Scheme Used

This manual has been prepared in outline form to facilitate cross-referencing and later updating. The identification scheme that is used employs Roman numerals, capital and small letters, and numbers in the sequence: I A 1 a (1) (a) [1] [a]. The major divisions, I, II, III, IV, etc., are called chapters. All other subdivisions are called "sections," with sections starting with capital letters (A, B, etc.) called "major sections." You are now, for example, reading Section I-C 1. To facilitate later updating, references within the manual are to sections and not pages.

2. Pagination

Each major section of this manual (e.g., I-C) starts on a new page, and pages are numbered within each major section. For example, this is Section I-C Page 1, or the first page of Section I-C.

3. Page Update Date

Immediately under each page number is the date that the page was drafted or revised. When a page has been revised, the date of the immediately previous version is also given in parentheses with the letter "s" meaning superseded." For example, III-B Page 1 dated 1 Jul 76 was revised on 8 Mar 85. The page number on the revised page would appear as:

III-B Page 1  
8 Mar 85  
(s. 1 Jul 76)

When updating the manual, new material that was not previously part of the text would not require the letter "s." For example, IV-E Page 6 originated on 8 Mar 85 would appear as:

IV-E Page 6  
8 Mar 85

4. Table and Figure Identification

Both tables and figures are numbered sequentially within a major section, with a hyphen before the table or figure number. Examples are: Table VIII-B-1, Table VIII-B-2, Figure VI-A-1.

D. Keeping This Manual Up to Date

1. Updated Pages Should be Inserted as Received

It is anticipated that sections of this manual will be periodically corrected, revised, or otherwise updated. New pages should be inserted as soon as they are received. This will not only keep the manual up to date, but will facilitate adding pages received at an even later date. Appropriate instructions covering which pages to add and delete will accompany distributed update pages. When it appears useful, a list will also be provided showing the page numbers and dates of all pages that should be in the manual at that time.

2. Request for Updates

To be placed on the distribution list to receive updates to this manual, write to:

Chief  
ARI Field Unit-Fort Hood  
HQ TCATA (PERI-OH)  
Fort Hood, Texas 76544-5065

E. Reporting Problems and Suggestions for Improvement

As previously noted, it is anticipated that this manual will periodically be updated to improve its utility. To report errors, problems, or suggestions, write to:

Chief  
ARI Field Unit-Fort Hood  
HQ TCATA (PERI-OH)  
Fort Hood, Texas 76544-5065



Chapter II: Major Questionnaire Types and Administration Procedures

A. Overview

This chapter briefly summarizes the different types of questionnaires discussed in this manual (Section II-B) and ways that questionnaires may be administered (Section II-C). Detailed guidelines regarding what to do in a given situation are included in subsequent chapters. Issues to consider when deciding whether to use a structured interview or some other type of questionnaire are presented in Section II-D, which also notes that combinations of methods may be employed. It is concluded that both structured interviews and other types of questionnaires have their place. Each has strengths and limitations which must be taken into account when identifying which instruments to use.

B. Types of Questionnaires Discussed in This Manual

There are a number of techniques of data collection that can be used to measure human attributes, attitudes, opinions, and behavior. Attitude and opinion are closely aligned if not overlapping. Opinions are restricted to verbalized attitudes. Attitudes are sometimes unconscious or nonverbalized. Some of the methods of data collection are observation, personal and public records, specific performances, sociology, interviews, questionnaires, rating scales, pictorial techniques, projective techniques, achievement testing, and psychological testing. For this manual, however, attention has been restricted to a more limited number of data collection techniques: certain paper-and-pencil types of instruments broadly classed as questionnaires as defined in Section I-A 2, and including only some of the techniques mentioned above. A distinction has also been made in this manual between open-ended questionnaire items and closed-end items. Open-ended items are those which permit respondents to express their opinions in their own words, and to indicate any qualifications they wish. The amount of freedom the respondent will be given in expressing an answer to an open-ended item is partly determined by the questionnaire designer. Closed-end items use response alternatives. Respondents are directed to select one or more of the response alternatives from a closed set. Closed-end items frequently used are multiple choice, true-false, checklist, rating scale, and forced-choice. Survey items have been roughly classified into two groups: open-ended items and closed-end items.

It is common to use interview surveys to ask questions and record answers. Structured interviews are included within the definition of questionnaires used, since typically an interview form is developed and used by an interviewer both for asking questions and recording responses, much like a self-administered questionnaire. On the other hand, the unstructured interview makes no use of structured data collection forms. The interviewers are permitted to discuss the subject matter as they see fit with no particular order or sequence. Of course, other interviews fall somewhere between these two extremes. In any case, unstructured interviews, where no structured response forms are used, are not included within the definition of questionnaires used in this manual.

C. Ways That Questionnaires Can Be Administered

There are a number of respects in which questionnaire administrations may vary. However, in the usual field test settings, the typical questionnaire administration situation involves paper-and-pencil materials with the author/test officer administering the questionnaire face-to-face with a group of test players or evaluators.

1. Group Versus Individual Administration

Given a printed questionnaire, calendar time is saved by group administration. Group administration allows the opportunity for a questionnaire administrator to explain the survey and answer questions about items. The task of statistical analysis can be initiated with less delay than if one were waiting on a series of individual administrations. An important determinant of group vs. individual is the time at which people complete their participation in the test. Most often all participants are through at the same time. All would be available for questionnaire administration as soon as they could be brought to an appropriate place or places. Prompt group administration gives the same short amount of time for forgetting about test events by those who become the respondents. Group administration generally has a high cooperation rate. If there is an administrator, his/her time is conserved directly in proportion to the number of respondents he/she has in each administrative session. An advantage of group administration is low cost.

2. Author-Administered Questionnaires

When the test officer or administrator who is familiar with the content of the questionnaire and the test's purposes/objectives can administer the questionnaire, some advantages can be gained. The administrator's instructions and appeals may increase the number of respondents having desirable motivation to complete the questionnaire by giving appropriate consideration to each item. If one employs a self-administration procedure, such as might occur in a mailed-out questionnaire, or if a poorly prepared stand-in plays the role of administrator, then the respondents must derive their instructions and some of their motivation from printed instructions (or from the poorly prepared stand-in). More things usually can end up going wrong when questionnaires are self-administered than when they are administered by a test administrator.

3. Remote Administrations

From the test officers' point of view, remote administration refers to a questionnaire administration event that they cannot conduct because of its distance from them and/or other demands on their time. This dimension, remote versus face-to-face, is similar but not identical to the previously noted dimension, self-administered versus author administered.

To avoid the possible disadvantages of self-administered questionnaires, the test officer must be able to afford another administrator, train him/her in the knowledge and skills associated with effective administration, and transport him/her to the "remote" administration location. If multiple administrations having location or timing differences which preclude the same administrator from handling them are required, it would appear that the chances are increased that more respondents will experience more "difficulties" in answering the questions. For this type of questionnaire administration, the questionnaire itself would require careful design associated with items and instructions.

4. Other Materiel Modes

Providing the respondents with a printed questionnaire form, and a pencil to mark/write their responses, is the most common questionnaire administration procedure in field evaluations. In addition, other presentation modes have been used. In a card-sorting procedure that has been used with individuals and groups, each respondent reads statements of candidate problems and then places the card into the appropriate pile according to his/her judgment of the severity of the "problem." Rarer because of expense and logistics problems is the setting up of a computer terminal where each respondent enters (types in) answers to questions that are displayed on a cathode ray tube (or other computer display device). Chapter XII presents many other considerations related to questionnaire administration.

D. Structured Interviews Versus Other Types of Questionnaires

1. Issues to Consider

When deciding whether to use a structured interview or another type of questionnaire, a number of issues should be considered.

Included are the following:

- a. To develop questionnaire items, a focus group may be interviewed. Their comments can be used to develop hypotheses and refine questions. This information can be adapted to an interview guide and interview items.
- b. Interview items should not use a dichotomous response set. Multiple choice and open-ended questions provide the opportunity for probing.
- c. If a structured interview is used, there must be enough qualified interviewers to expeditiously process all interviewees. Sometimes there are only a few personnel to be interviewed, or there is plenty of time available for interviews, so only one or two interviewers will be necessary. In other situations, maybe only an hour or so may be available per interviewee; in these cases, a large number of qualified interviewers must be available.
- d. Face-to-face interviews have a higher response rate than mail surveys.
- e. In most cases, respondents have a greater tendency to answer open-ended questions in an interview than when response is by paper and pencil.
- f. It is possible to adapt face-to-face interview guides for telephone surveys. Oral labeling of the scale points should be assessed on a pilot survey to be sure that the responses are not biased by the oral presentation of the scale.
- g. Telephone interviews are faster to perform than mail surveys.
- h. Interviews conducted by telephone require an interview structure that promotes a high interaction between the interviewer and respondent.
- i. Group-administered paper-and-pencil questionnaires may be less expensive, more anonymous, and completed faster than the same number of interviews.
- j. Respondents seem to be less likely to report unfavorable things in an interview than in an anonymous questionnaire. Typically, questionnaires are also more likely than interviews to produce self-revealing data.

- k. Issues involving socially acceptable or unacceptable attitudes and behaviors will elicit more response bias.
- l. During interviews, respondents often have a tendency to try to support the norms that they assume the interviewer adheres to.
- m. Interviewers with biases on the issues under discussion may reflect them in the content they record, as well as in what they fail to record.
- n. Ethnic background differences between interviewer and respondent probably will not influence the survey results unless the items have a racial content or are found to be threatening.
- o. Although a structured interview using open-ended questions may produce more complete information than a typical questionnaire containing the same questions, empirical research seems to indicate that responses to the typical questionnaire are more reliable; i.e., more consistent. Structured interviews using closed-end questions appear to be as reliable as paper-and-pencil questionnaires.
- p. It may be difficult to code a combination of open-ended and closed-end items for interview surveys. (See Section XIII-B, Scoring Questionnaire Responses.)

## 2. Combinations of Methods

There are some situations where a combination of methods of questioning might be used:

- a. An interview might be used to obtain information for designing a paper-and-pencil questionnaire.
- b. Personal interviews or telephone interviews might be used for respondents who do not return questionnaires administered remotely (such as mail questionnaires).
- c. When respondents are unable to give complete information during an interview, they can be left a copy of a questionnaire to complete and mail in, so that the necessity for a call-back is eliminated.

## 3. Conclusion

Both structured interviews and other types of questionnaires appear to have their advantages and disadvantages. The choice of which to use may well depend upon costs, which are generally lower for the typical questionnaire. The typical questionnaire is apparently more reliable, while the structured interview may provide more unique and more abundant information. If the dimensions of a problem have not been explored before, the best compromise would appear to be to use the interview approach with open-ended items to uncover the dimensions, and follow this by the use of the paper-and-pencil questionnaire with closed-end items to obtain more specific information.

Chapter III: Content of Questionnaire Items

A. Overview

The recommended general steps in preparing a questionnaire include preliminary planning, determining the content of questionnaire items, selecting question forms, wording of questions, formulating the questionnaire, and pretesting. As part of preliminary planning, the information required has to be determined, as do procedures required for administration, sample size, location, frequency of administration, experimental design of the field test, and analyses to be used. Selecting question forms is a function of the content of the questionnaire items and requires knowledge of types of questionnaire items and scaling techniques. The wording of questions is the most critical and most difficult step. Formulating the questionnaire includes formatting, sequencing of questions, consideration of data reduction and analysis techniques, determining basic data needed, and insuring adequate coverage of required field test data. Pretesting involves using a small but representative group to insure that all questions are understandable and unambiguous.

This chapter considers the content of questionnaire items. Methods for determining questionnaire content are discussed first, and then other considerations related to questionnaire content are presented. The other steps noted above are discussed in subsequent chapters.

B. Determining Questionnaire Content Preliminary Research

1. Preliminary Research

If you have the job of developing a questionnaire for a field test, there are several things that should be done before starting to write questionnaire items.

- a. Learn the test's objectives and issues. Read the Outline Test Plan in order to learn what it says the test's purpose, scope, and objectives are. All data collection effort, including questionnaire administration, should be consistent with and supportive of the test's objectives. Read the Independent Evaluation Plan, with its discussion of issues and of ways of collecting data on the issues.
- b. What performance measures are planned for the test? One may be fortunate enough to be involved with a test for which the Detailed Test Plan has to a large extent been written. Try to discover what performance measures/data are to be collected. If performance data is to be collected on some aspects of the functioning of the system to be tested, then it may not be necessary to assess these functions via questionnaire items. Make a list of what should be measured to meet the objectives of the field test. The list will include variables that are configured into categories. The list should not include any questions.
- c. Consult others and prior test plans and reports. Many tests at CDEC and TCATA (and elsewhere) follow-up, or are similar to, prior testing. As a consequence, information may be readily available regarding prior related or similar tests. Test files or the Technical Information Center may provide a source for obtaining test plans and reports on relevant prior tests conducted by Army field test/experimentation agencies.
- d. Consult others and develop an analysis plan. The Technical Information Center may provide guidance for data analysis. Develop an analysis plan with a list of variables to be measured. The analysis plan identifies dependent and independent variables. It also identifies which variables to control and any intervening variables.

Preliminary research requires an understanding of the objectives of the test plan, a list of the variables to be measured, and a plan for analysis of the data.

2. Using Interviews to Determine Questionnaire Content

If one's degree of experience seems meager relative to the complexities of the evaluation problem, he/she may employ group and/or individual interviews to assist in determining questionnaire content. Preferably, this would be done after taking the steps noted above. The less one knows about a subject, the less structure one can impose on an interview dealing with the subject.

- a. Conducting an unstructured group interview. Personnel are needed who have relevant operating experience with the system to be tested/evaluated - or with a sufficiently similar system. Arrange a common meeting place and time with about five to ten of them. It would be advantageous to have a meeting place that was not cramped for space, had comfortable chairs, a comfortable temperature, and where all discussants were free from other sources of distraction (sights and sounds, mainly).

If the interviewer's age and rank are several steps above or below the age and rank of the members of a homogeneous group of discussants, try (before the meeting) to get a person who is their contemporary (peer) in age and rank to lead and coordinate the discussions. Why? Because a mismatch may inhibit their discussion or produce too much submissive, agreeing behavior on their part.

If notes are being taken or the discussion is being tape recorded, one should be unobtrusive about it. Don't shove/point a microphone at people as they start to speak. They may be inhibited by this, or they may become "hams."

The first several minutes should be spent in establishing rapport with the group. The purpose of the session should be covered, introduction of group members made, and other warm-up devices used. The objective is to motivate as many respondents to give comments as possible. In the remainder of the session, any or all of the following information-eliciting devices could be used:

- (1) Discuss samples of the control item--ask the general question: "What problems have you had with this piece of equipment or system?" Follow up with who, what, where, when and why. Attempt to maximize the number of potential or actual problems posed. Strive for clarification of problem ideas, but do not criticize the comments, even if they are redundant with a previous contribution by the respondent or other respondents.
- (2) Ask: "What do you consider to be the most important features (characteristics, qualities, etc.) of this equipment or system when used in the field?" Strive to get a multitude of adjectives and phrases here (e.g., ease of operation, weight, durability, portability, etc.).

- (3) Use the aided recall technique: "Can you remember where and when you have encountered problems with this system?" (e.g., at night; when it's damp, etc.).
- (4) The way survey issues are discussed will help in selecting vocabulary and phrasing questions.
- (5) 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. Development of questionnaire items is based on hypotheses that have been developed. The hypotheses are presented to a group of individuals who are subject matter experts, and they perform a preliminary assessment of the hypotheses. The questionnaire may require modification if the hypotheses are not viable.

The recorded comments should be categorized and arranged by frequency. For example, how many of the comments on system operation stressed failure considerations?

- b. Conduct semistructured personal interviews. Information produced from the unstructured group interviews provides general guidance to the specific evaluative information desired. As a next step, or as an alternative step to the group interview, one may employ a small number of representative respondents in a person-to-person interview format.

In this method of interviewing, the interviewers are given only general instructions on the type of information desired. They are left free to ask the necessary direct questions to obtain this information, using the wording and the order that seems most appropriate in the context of each interview. These interviews, like the unstructured group sessions, are useful in obtaining a clearer understanding of problems, and in determining what areas (evaluation criteria) should be included on the pilot questionnaire.

The only structure to the semistructured interview comes from a set of question categories that must be raised sometime during the interview. Questions on system experience, positive and negative features, and problems in field use, for example, can be phrased in any manner or sequence. Probing questions of the type: "Why do you feel that way?," "What do you mean by that statement?," and "What other reasons do you have?" can be utilized until the interviewers are satisfied that they have the necessary information considering time limitations, data requirements, and the willingness and ability of the respondents to verbalize their views. Interview forms should be designed to allow the interviewer sufficient space for writing notes and comments.

In the semistructured interview, the interviewer has some flexibility in formulating and asking questions. This technique can, therefore, be only as effective in obtaining complete, objective, and unbiased information as the interviewer is skilled in formulating and asking questions. Thus, interviewers may have to be trained in using this technique.

When interviews are used as the basis for a future questionnaire, the questions need to be carefully stated so that they are eliciting data which will enable the interviewer to construct questions which address the stated objectives and issues of the research. Once the questionnaire items have been identified, the items need to be assembled into a logical sequence. They then need to be administered to a sample of respondents who have a background similar to the audience to which the questionnaire was originally targeted. Information obtained from the sample administration is used to refine questionnaire items.

- c. Develop the questionnaire. In the development phase of a questionnaire, an open-ended response format can be useful in selecting meaningful response alternatives for a multiple choice format. Open-ended questions administered to a sample of the target population will provide responses that can then be phrased in the spontaneous wording of the individuals in the sample. The questionnaire items can be pretested using an open-ended response format on respondents who are representative of the eventual test population. Prior to pretesting the open-ended questions, the test officer needs to be sensitive to the phrasing of the questions since inadvertent phrasing of the open-ended questions can sometimes modify responses in unrecognized and unintended ways. The use of open-ended response formats and interviews should enable the formulation of a questionnaire to obtain evaluative information. These interviews will provide guidance to the formulation of a sound survey instrument in the following respects:

- (1) A better understanding of the factors or criteria which make up the mental set of individuals in evaluating systems and equipment.
- (2) Some idea of the range of favorable and unfavorable opinions toward the system for each factor.
- (3) Tentative knowledge of individual and group differential opinions toward the system tested.

Therefore, before drafting the pretest questionnaire, the researcher must have a feel for: question categories (e.g., problem areas, positive aspects); response categories (e.g., evaluative factors); and the type of system operations information which is needed (e.g., In evaluating a new helmet suspension system, does respondent wear eyeglasses?).

3. Using the Critical Incident Technique to Determine Questionnaire Content

The critical incident technique consists of a set of procedures for collecting direct observations of human behavior in such a way as to facilitate their potential usefulness either in solving practical problems or in developing broad psychological principles. The technique calls for collecting observed incidents of behavior that have special significance and meet systematically defined criteria. It can be of assistance, therefore, in helping to determine the content of items to be included in a questionnaire.

Although there are a number of variations in the critical incident technique, the basic procedure consists of collecting records of specific behaviors related to the topic of concern. The behaviors might be noted by observers, or individuals can be asked to recall and record past specific behaviors judged to provide significant or critical evidence related to the topic of concern. As appropriate, behaviors related both positively and negatively to the area of concern should be noted. The records of behavior that are collected can then be analyzed and used as a basis for determining questionnaire content.

One of the examples of the use of the critical incident technique reported by Flanagan in the articles noted in Section III-B 3, had to do with a study of combat leadership in the United States Army Air Forces in 1944. It represented "the first large-scale, systematic effort to gather specific incidents of effective or ineffective behavior with respect to a designated activity. The instructions asked the combat veterans to report incidents observed by them that involved behavior which was especially helpful or especially inadequate in accomplishing the assigned mission. The statement finished with the request, 'Describe the officer's action. What did he do?' Several thousand incidents were collected in this way and analyzed to provide a relatively objective and factual definition of combat leadership. The resulting set of descriptive categories was called the 'critical requirements' of combat leadership" (p. 328).

For more information on the critical incident technique, see, for example, the following two sources:

- a. Barnes, T. I. (1960). The critical incident technique. Sociology and Social Research, 44, 345-347.
- b. Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-358.

4. Using Impressions of a Topic to Determine Attitude Scale Content

When the questionnaire is an attitude scale, a useful method for selecting items for it is to ask a group of individuals to write six statements giving their impressions of a topic, such as Army pay. From these, some smaller number of statements can be selected that are readable, intelligible, and capable of classification. These statements can then be sorted into several categories, such as the status of the topic and its good and bad features.

C. Other Considerations Related to Questionnaire Content

This section discusses a number of topics related to questionnaire content: questions that should be asked related to questionnaire content; sources of bias in questionnaire construction; and characteristics of good questions that affect questionnaire content.

1. Questions That Should Be Asked Related to Questionnaire Content

Asking yourself the following five questions may lay the foundation for a far more valuable questionnaire than would otherwise be produced. If you can't answer these questions, be sure to read or re-read the Outline Test Plan and the Independent Evaluation Plan.

- a. Who needs the information? Knowledge of who needs the information will provide a source in the event answers are needed to the following four questions.
- b. What decisions will be made based on your information? This will tell in part why the information is needed. Depending on what decision is going to be made, some kinds of information will make a difference and should be collected, and other kinds will not.

Suppose, for example, information is to be collected as a part of a test comparing a new item of equipment with an old standard item. The nature of the decision to be made is clear enough. It will be either selection of the new equipment, or retention of the old with which it is being compared. The basis for the decision will usually also be clear from the small development requirement (SDR) or qualitative materiel requirement (QMR) which led to the development of the item being tested. Analysis of the QMR will identify the qualitative requirements the new equipment must have, and will give the start needed to develop questions.

- c. What facts will affect the decision? While this may be a difficult question to answer, trying to do so should identify items of information that should be sought with the questionnaire. It may also head off the collection of unnecessary information.
- d. Whom are you asking? To get good information, not only must a good question be asked, but it must be asked of someone who has the answer. It would not, for example, be reasonable to ask support troops in a supply depot questions about combat operations.

- e. What are the consequences of a wrong answer? While this basically is an administrative question, it has an important bearing on field questionnaire design. Clearly, if it makes little difference which of two alternatives is chosen, it makes little difference if the information is collected. On the other hand, if there is a chance that substantial dollar savings will result from the use of a more effective training technique, or that millions of dollars will be wasted by buying a new piece of equipment which is not better than the old, it is necessary to design tests very well, and ask the right questions with great care.

## 2. Refining Questions

Early versions of questions usually need to be refined. The following approaches will assist in developing better questions:

- a. Try out questions on co-workers.
- b. Identify problems in question wording prior to pretesting.
- c. Pretest the questionnaire, and modify as needed. This should help in making the questionnaire easier for the respondents to use, and to assure meeting the objectives of the field test.

## 3. Sources of Bias in Questionnaire Construction

Two primary sources of bias in questionnaire construction that have been identified are investigator bias and question bias.

- a. Investigator bias arises from: choice of subject matter; study design and procedure; unfair or loaded phrasing of questions; and interpretation and reporting of results. Sources of such biases include: the questionnaire developers' relationships with the clients; their personal involvement in a particular theoretical position or research technique; and those personal traits attributable to class, race, or political ideology. To reduce the impact of such bias, questionnaire developers need to: be aware of the problems; seek critiques from independent sources; carefully review previously published related reports; and continue pursuing technical improvement in their investigations.
- b. Four ways that have been suggested of minimizing question bias when asking opinion questions are: ask many questions on the same topic; determine by scale analysis whether questions ask the respondents about the same dimensions of opinion (see Chapter V); ask "How strongly do you feel about this?" after each opinion question; and relate the content of opinion to the intensity of feeling.



3

## Chapter IV: Types of Questionnaire Items

### A. Overview

This chapter discusses various types of questionnaire items: open-ended items (Section IV-B), multiple choice items (Section IV-C), rating scale items (Section IV-D), behavioral scale items (Section IV-E), ranking items (Section IV-F), forced choice and paired-comparison items (Section IV-G), card sorting items/tasks (Section IV-H), and semantic differential items (Section IV-I). For each of these major item types, definitions and examples are presented, advantages and disadvantages are noted, and recommendations regarding their use in Army field test evaluations are given. Other types of items are noted in Section IV-J: checklists, matching items, arrangement items, and formats providing for supplementary responses.

It may be noted that a number of ways have been utilized in the professional literature for differentiating and classifying item types. Which types are special cases of other types could be debated at length. Unanimous agreement with the definitions given in this manual cannot, therefore, be anticipated.

B. Open-Ended Items

1. Definition and Examples

Open-ended items are those which permit respondents to express their answers to the questions in their own words, and to indicate any qualifications they wish. They are like general questions asked in an unstructured interview. By contrast, in a closed-end item, all the answers/choices/responses permitted are displayed, and respondents need only to check their preferred choices. Examples of open-ended items are shown in Figure IV-B-1.

Figure IV-B-1

Examples of Open-Ended Items

1. Describe any problems you experienced in moving through the test course while wearing the new PRC-99 radio harness.

\_\_\_\_\_

2. The M16 rifle is:

\_\_\_\_\_

3. What do you think of the AR-15 rifle sight?

\_\_\_\_\_

2. Advantages of Open-Ended Items

- a. Questions with open-ended response formats allow the respondents considerable latitude in their responses.
- b. Open-ended items allow for the expression of middle opinions that closed-end items with two choices would not.
- c. Open-ended items allow for the expression of issues of concern that may not have been identified by the question writer.
- d. Open-ended items allow researchers to obtain answers that are unanticipated; unique information may be provided.
- e. Open-ended items are very easy to ask. This is useful when the question writer either does not know, or is not certain about, the range of possible alternative answers.

- f. With an open-ended question, it is possible to find out what is salient to the respondents, what their frame of reference is, and how strongly they feel.
- g. Open-ended questions permit respondents to describe more closely and fully their real views.
- h. There are times when more valid answers may be obtained from open-ended than closed-end items. For example, there may be a tendency for respondents to inflate yearly income figures. Providing response alternatives may result in an even greater inflation.
- i. Answers to open-ended questions may be useful when treated as anecdotal material.
- j. Respondents like the opportunity to answer some questions in their own words.

### 3. Disadvantages of Open-Ended Items

- a. Open-ended items are time consuming for the respondent.
- b. Open-ended questions which are self-administered and/or group-administered place a burden on the reading and writing skills of the respondent.
- c. Asking people to answer questions in their own words increases the task difficulty, and can affect the rate of response. For example, respondents may say that they have no problems rather than taking the time to write out what the problems are. Item 1 in Figure IV-B-1 is poor in this respect, but item 2 is worse.
- d. Only highly motivated respondents will take the time to write a complete answer to each question.
- e. Open-ended items often leave the respondents on their own to determine what is relevant in the evaluation. For instance, item 2 in Figure IV-B-1 leaves the respondents to determine what is relevant in evaluating the M16 rifle. This is inappropriate. Open-ended questions should not be used to bypass the understanding of operations that the questionnaire writer should have or should acquire before preparing the final version of the questionnaire.
- f. Questionnaires that use closed-end items are generally more reliable than those using open-ended items.

- g. Open-ended questions, answered by motivated respondents, are capable of overloading data analysts. They usually cannot be handled by machine analysis methods without lengthy preliminary steps. Analysis of the responses to an open-ended question usually must be done by someone who has substantial knowledge about the question's content, rather than by a statistical clerk. They are often difficult to code for analyses. Thus, the data analysis task can grow into a major project and problem.
- h. Open-ended questions may be easier to misinterpret since the respondent does not have a set of response alternatives available which might in themselves provide the proper frame of reference.
- i. Much of the material obtained from an open-ended question may be repetitious or irrelevant.
- j. Since open-ended questions are more time consuming, a constraint is placed on the number of questions that can be asked.
- k. Open-ended questions are more subject to interviewer variations than are closed-end questions.
- l. Open-ended items are often harder for the respondent to answer than closed-end questions. For example, respondents, when asked their annual income, may have to struggle to come up with relatively specific figures, whereas when response alternatives are presented, they need only indicate one of a number of ranges of income.
- m. Inadvertent phrasing of open-ended questions can sometimes modify responses in unrecognized and unintended ways. It is difficult to predict in advance which words will bias an item. Subtle words appear to cause more distortion than blatantly biasing words.

#### 4. Recommendations Regarding Use

- a. Open-ended questions should be rarely used and, even then, such questions should sharply focus respondents' attention and thereby reduce their writing burden.
- b. Closed questions are better for self-administered questionnaires than open questions.
- c. In situations where time and money constraints are paramount, it would be more appropriate to use closed questions.

- d. Closed questions are preferred for surveys where the responses would more likely be dichotomous.
- e. For collecting nominal data, the researcher has a choice about whether to ask open-ended or closed-end questions.
- f. When responses can be obtained by degree (for example, strongly agree to strongly disagree), a closed-end question would be superior to an open-ended question.
- g. Sometimes a good procedure is to use an open-ended question with a small number of respondents as a pretest, in order to find out what the range of alternatives is. It may then be possible to construct good closed-end questions that will be faster to administer and easier to analyze.
- h. Open-ended questions are most useful when there are too many possible responses to be listed or foreseen; when it is important to measure the saliency of an issue to the respondent; or when a rapport-building device is needed in an interview.
- i. To obtain in-depth information on various content areas, a more focused and guided approach would be the use of an interview with open questions.
- j. Use long open questions with familiar wording for questions with potentially threatening content.
- k. It is sometimes useful to include one or more open-ended questions along with closed-end questions in order to obtain verbatim responses or comments that can be used to provide "flavor" of responses in a report.

C. Multiple Choice Items

1. Definition and Examples

In a multiple choice item, the respondent's task is to choose the appropriate or best answer from several given answers or options. As used here, multiple choice items include dichotomous or two-choice items as special cases. And, since only the permitted answers are available for selection, the multiple choice item may also be termed a closed-end item.

Examples of multiple choice items are shown in Figure IV-C-1. Items 3, 4, and 5 are dichotomous, i.e., provide two response alternatives.

A comparison of true-false items with nondichotomous multiple choice items is made in Section VI-G, since they are issues related to the number of response alternatives.

2. Advantages of Multiple Choice Items

- a. As seen in item 2 of Figure IV-C-1, the questionnaire writer may select different numbers of response alternatives depending upon knowledge of the respondent's experience or depending upon the decision to allow or disallow respondents to "sit on the fence" by including a "no preference" alternative. (See Section VI-C for wording of items, and Section VI-G regarding the number of response alternatives to employ.)
- b. Responses are more reliable when response alternatives are provided for respondents.
- c. Interpretation of responses is more reliable when response alternatives are provided to respondents.
- d. Dichotomous items are relatively easy to develop, and permit rapid analyses.
- e. Complex questions can often be broken down into two or more simpler questions.
- f. Multiple choice items are easily scored, which means that data analysis is a relatively inexpensive process requiring no special content expertise.
- g. Multiple choice items require considerably less time per respondent answer than open-ended items.
- h. Multiple choice items put all persons on the same footing when answering. That is, each person will be able to consider the same range of alternatives when choosing an answer.
- i. Multiple choice items are easy to administer.

Figure IV-C-1

Examples of Multiple Choice Items

1. What do you consider the most important characteristic of a good helmet? (Check one)  
 Comfort  
 Stability  
 Utility for wash basin  
 Protection  
 Weight
2. Which do you prefer, the M16 or the M14 rifle? (Check one)  
 M14  
 M16  
 No preference
3. Were you able to fire effectively from the frontal parapet emplacement?  
 Yes     No
4. Which do you prefer, the ABC helmet or the XYZ helmet?  
 ABC helmet     XYZ helmet
5. The M16 is a better rifle than the M14.  
 True     False
6. What is your marital status?  
 Single  
 Married  
 Divorced  
 Other (e.g., separated, widowed, etc.)

3. Disadvantages of Multiple Choice Items

- a. Dichotomous items force the respondents to make a choice even though they may feel there are no differences between the alternatives, or they do not know enough about either to validly choose one. Furthermore, respondents are not permitted to say how much better one alternative is than the other.
- b. Two alternatives might not be enough for some types of questions. The question designer may oversimplify an issue by forcing it into two categories.
- c. There may be a tendency for respondents to choose an answer on the basis of a response set. (See Chapter XII.)
- d. Unless care is taken in the construction of multiple choice items, the response alternatives may overlap.
- e. The question maker has to know the full range of significant possible alternatives at the time the multiple choice question is formulated.
- f. Multiple choice items must be worded with very great care. Otherwise, the information obtained may not be valid.
- g. With dichotomous items, any slight language difficulty or misunderstanding of even one word could change the answer from one extreme to another.

4. Recommendations Regarding Use

- a. For some purposes, the dichotomous question (two response alternatives) may be an improvement over the open-ended question in that it provides for faster and more economical analysis of data. However, it requires more care in its development.
- b. Generally speaking, dichotomous multiple choice questions should be avoided. If used, they should probably be followed-up to determine the reason for a given response.
- c. Nondichotomous multiple choice items are popular and have wide utility. They are recommended for general use as appropriate.
- d. Forced response and multiple choice items are desired when measuring soft data such as opinions. Checklists are recommended for hard data such as physical aspects of a job analysis or a broad generalization for measuring opinions prior to a later survey.

- e. The development of questionnaire items should include pilot testing using open-ended items which are later converted to multiple choice items.
- f. No one scaling format has consistently been superior to another. Rating scales need to be evaluated on other criteria than number of scale points, vertical and horizontal formats, and unipolar or bipolar scales.
- g. Prior to multiple choice format selection, the type of measurement scale and data analysis should be identified.
- h. Multiple choice items represent measurement scales which are nominal, ordinal, or interval. These measurement categories indicate the rules for assigning numbers to the data so that the appropriate statistical analyses can be performed.
- i. Ordinal measurement scales are common in surveys where respondents are required to rank items or to use a paired-comparison method.
- j. One item cannot adequately cover a topic area. It is necessary to develop many items to avoid obtaining only surface facts, and to provide the researcher with a deeper understanding of the relevant experience of the respondents.
- k. Multiple choice items can be developed which measure higher order objectives.
- l. If multiple questions are asked about different possible responses to a problem, separate specific questions that can be understood by all respondents and easily interpreted are required.
- m. The length of an item may possibly modify the response style. Researchers may wish to develop alternate versions of questionnaire items where the different versions are of different lengths. This would allow comparison of the effect of item length on responses.

D. Rating Scale Items

1. Definitions and Examples

Rating scale items are a variation of multiple choice items. They are a means of assigning a numerical value to a person's judgment about some object. They call for the assignment of responses either along an unbroken continuum or in ordered categories along the continuum. The end result is the attachment of numbers to those assignments. Ratings may be made concerning almost anything, including people, groups, ourselves, objects, and systems.

There are a number of different forms of rating scale items, only two of which are shown here. Figure IV-D-1 shows examples of "numerical" scales. In item 1, a sequence of defined numbers is provided for the respondent.

Figure IV-D-1

Examples of Numerical Rating Scale Items

1. The cleaning kit for the M16 rifle is

- \_\_\_\_\_ 7 very easy to use.
- \_\_\_\_\_ 6 quite easy to use.
- \_\_\_\_\_ 5 fairly easy to use.
- \_\_\_\_\_ 4 borderline.
- \_\_\_\_\_ 3 fairly difficult to use.
- \_\_\_\_\_ 2 quite difficult to use.
- \_\_\_\_\_ 1 very difficult to use.

2. How satisfied or dissatisfied are you with the type of furniture in the barracks?

- \_\_\_\_\_ Very satisfied
- \_\_\_\_\_ Satisfied
- \_\_\_\_\_ Borderline
- \_\_\_\_\_ Dissatisfied
- \_\_\_\_\_ Very dissatisfied

3. The training that I have received at Fort Hood has been

- \_\_\_\_\_ very challenging.
- \_\_\_\_\_ challenging.
- \_\_\_\_\_ borderline.
- \_\_\_\_\_ unchallenging.
- \_\_\_\_\_ very unchallenging.

The respondents are to indicate which defined number best fits their judgment about the object to be rated. Sometimes, the numbers are not shown on the form used by the respondent (e.g., items 2 and 3). Instead, the respondent reports in terms of descriptive cues and the numbers are attached later during analysis. The numbers assigned are in an arithmetic sequence, such as 5, 4, 3, 2, 1, depending upon the number of response alternatives used. They are usually assigned arbitrarily unless the response alternatives have been scaled using one of the procedures described in Section V-B. The order of perceived favorableness of commonly used words and phrases is discussed in Chapter VIII.

Figure IV-D-2 shows an example of a graphic rating scale. In the graphic scale, the descriptors are associated with points on a line or graph, and the respondent indicates a judgment by marking the point on the line which best fits the rating of the object. The line can be either horizontal or vertical. The graphic scale allows the respondent to place a judgment any place on the line. Thus, the respondents are not confined to discrete categories as they are with the numerical scale. It is, however, more difficult to score, but this can be facilitated with a stencil which divides the line into segments to which numbers are assigned.

The number of response alternatives to use is discussed in Section VI-G, the order of response alternatives in Section VI-H, and response anchoring in Chapter VII.

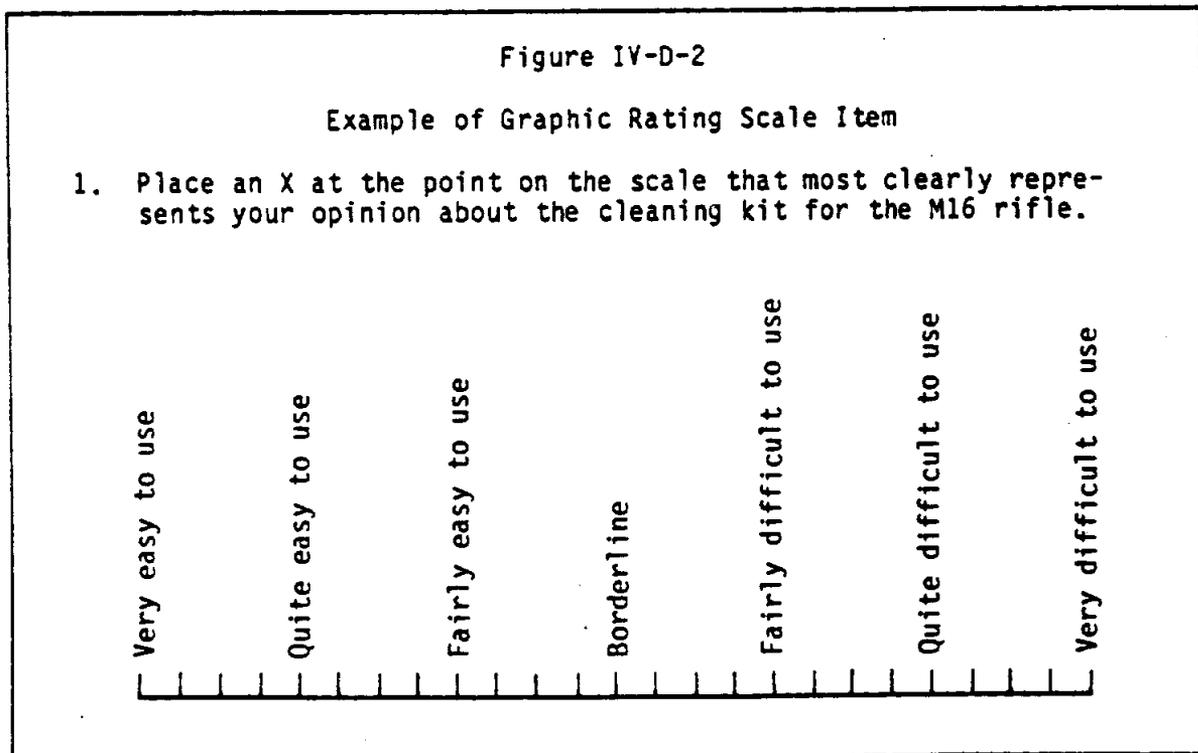
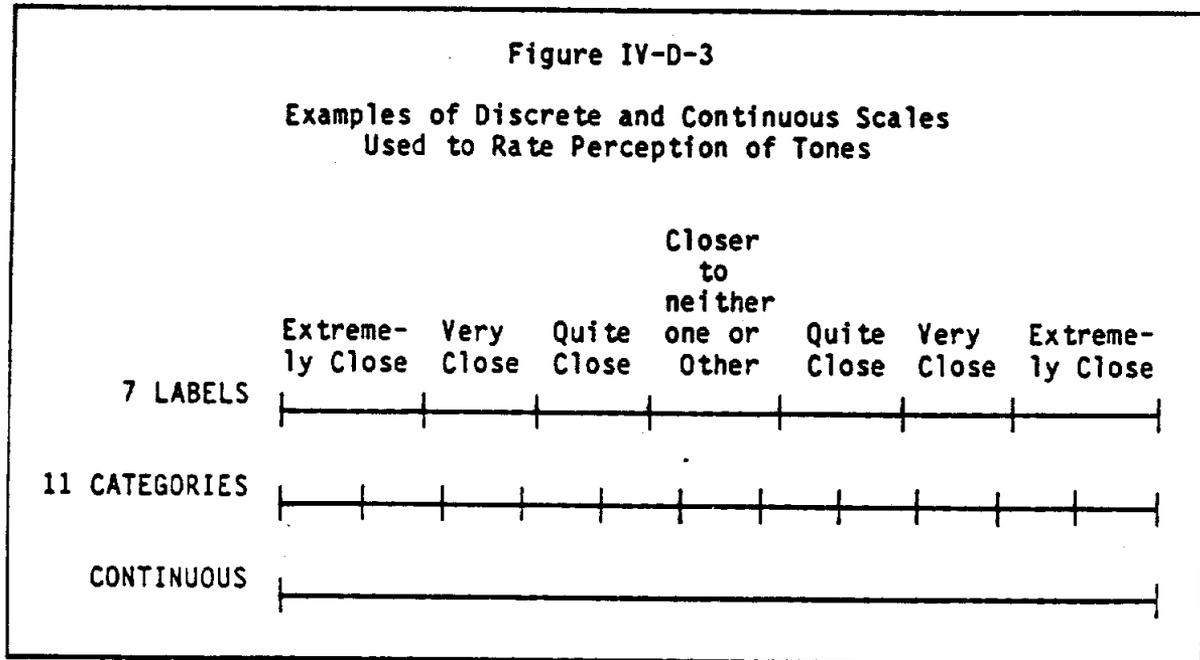


Figure IV-D-3 shows examples of continuous scales.

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. Continuous scales have been used in ergonomics to rate perception of a thermal stimulus as well as to rate perception of tones.



## 2. Advantages of Rating Scale Items

- a. When properly constructed, the rating scale reflects both the direction and degree of attitude or opinion, and the results are amenable to analysis using conventional statistical procedures.
- b. Graphic rating scales allow for as fine a discrimination as the respondent is capable of giving, and the fineness of scoring can be as great as desired.
- c. Rating scale items usually take less time to answer than do other types of items.
- d. Rating scale items can be applied to almost anything.
- e. Continuous scales may at times yield greater discrimination by raters.
- f. Rating scale items are generally more reliable than dichotomous multiple choice items. They may be more reliable than paired-comparison items.
- g. Manipulation of the anchors does not appear to greatly affect the results. The inadvertent use of mismatching antonyms with partial antonyms to anchor a rating scale may not jeopardize the reliability of the scale.

## 3. Disadvantages of Rating Scale Items

- a. Rating scale items are more vulnerable to biases and errors than other types of items such as forced choice items.
- b. Graphic rating scales are harder to score than other types of items. With a graphic scale item format, the verbal anchors are associated with points on a line, and the respondents indicate their judgment by marking the point on the line which best represents their judgment. Considerable effort and time are required to measure the pencil mark's exact location to the nearest portion of the line.
- c. The results obtained from the use of graphic rating scale items may imply a degree of precision/accuracy which is unwarranted.

4. Recommendations Regarding Use

- a. The use of rating scale items is highly recommended for most questionnaires.
- b. Rating scales present the sentence (stem) first, and require the respondent to select a response alternative to complete the sentence. The stem is supposed to be neutral so that the response alternatives contain different combinations of directionality (positive or negative) and intensity.
- c. Scales having apparently equal intervals should be employed. The respondent will assume or perceive that the distances between adjacent scale points are equal.
- d. Numbers can be presented along with verbal anchors.
- e. Applications which require greater discrimination could use scales with more than five or six categories, or with continuous lines.
- f. 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.
- g. Minor violations in the technique of scale development for bipolar anchors, such as quasi-polar anchors and phrases for anchors, do not appear to threaten the reliability of the instrument. Therefore, it is possible to establish new versions for bipolar anchors.

E. Behavioral Scale Items

1. Definition and Examples

Behavioral scale items are derived from the compilation of critical incidents (whether really critical or not). They were developed to encourage raters to observe behavior more accurately. Behavioral scales have evolved using different developmental procedures with divergent scaling foundations associated with Likert, Thurstone, and Guttman scales. There are a variety of behavioral scales such as Behaviorally Anchored Rating Scales (BARS), Behavioral Expectation Scales (BES), Behavioral Observation Scales (BOS), and Mixed Standard Scales (MSS).

Behavioral scales have customarily been used to evaluate individual performance on the job. There have been other applications that include assessing morale, and a tool to make decisions about the effectiveness of maintenance trainer equipment and actual equipment training.

Even though developmental procedures vary according to the type of behavioral scale, there are some commonalities. Behavioral scales are built on large numbers (in the hundreds) of critical incidents which are reduced in number by being fitted into performance dimensions and/or categories. There must be a specified level of agreement (usually somewhere between 60% and 80%) to retain a critical incident for inclusion in the scale. The critical incidents are anchored to the scale. Critical incidents describe a continuum of effective and ineffective behavior.

Procedures for constructing behavioral scale items, and evaluative comments about them, can be found in a number of sources including the following:

- a. 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.
- b. Borman, W. C. (1979). Format and training effects on rater accuracy and rater errors. Journal of Applied Psychology, 64, 410-421.
- c. 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)
- d. 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.
- e. Landy, F. J., & Barnes, J. L. (1979). Scaling behavioral anchors. Applied Psychological Measurement, 3(2), 193-200.
- f. 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.
- g. Motowidlo, S. J., & Borman, W. C. (1977). Behaviorally anchored scales for measuring morale in military units. Journal of Applied Psychology, 62(2), 177-183.
- h. 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 Arts and Sciences (MMAS) thesis prepared at U.S. Army Command and General Staff College, Fort Leavenworth, KS. (DTIC No. AD A097694)

Examples of behavioral scale items and dimensions are shown for BARS, BES, BOS, and MSS in Figures IV-E-1 through IV-E-4.

Figure IV-E-1

Examples of BARS's Seven Dimensions  
Describing Technician Behavior

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 technicians are well prepared when they arrive on the job, carry out maintenance procedures completely and thoroughly, and recognize and attend 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 technicians thoroughly understand system operation allowing them 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 technicians understand the systems that are interconnected with their 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.

From Wienclaw, R. A., & Hines, F. E. (1982, November). A model for determining cost and training effectiveness trade-offs. Training Equipment Interservice/Industry Training Equipment Conference, 405-416.

Figure IV-E-2

Example of BARS Items Representing  
Performance and Effort on the Job

<u>Scale Point</u>	<u>Behavioral Anchor</u>
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.

From Motowidlo, S. J., & Borman, W. C. (1977). Behaviorally anchored scales for measuring morale in military units. Journal of Applied Psychology, 62(2), 177-183.

Figure IV-E-3

Example of BOS Item Representing  
Description of Foreman's Job

Tells crew to inform him immediately of any unsafe condition.

Almost Never 1 2 3 4 5 Almost Always

From 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.

Figure IV-E-4

Example of MSS Items Representing  
Highway Patrol Stopping Vehicles for 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.

From Rosinger, G., Myers, L. B., Levy, G., Loar, M., Mohrman, S. A., & Stock, R. (1982). Development of behaviorally based performance appraisal system. Personnel Psychology, 35, 75-88.

2. Advantages of Behavioral Scale Items

- a. Raters may not be cognitively prepared to summarize and abstract accurately. More reliable ratings may be obtained on behavioral scales by using the jargon of raters, and by having raters maintain observational diaries.
- b. It has been found that it is possible to generalize a Behaviorally Anchored Rating Scale (BARS) instrument for use with similar populations in other organizations where the same types of tasks are being performed.
- c. Behavioral Expectation Scales (BES) can be used to clarify organizational policy, provide feedback, assess and improve individual performance, and identify divergent perceptions.
- d. Training programs of three hours and longer have the potential to increase rater accuracy.
- e. In situations where there is concern about halo and leniency errors, Mixed Standard Scales (MSS) would be appropriate to use if the developmental procedures are thorough.

3. Disadvantages of Behavioral Scale Items

- a. The time and effort involved in developing behavioral scale items may not be worth the investment unless there are other spin-offs for the use of this type of scale.
- b. Behavioral scales require quantification of items using a sample size of several hundred people; they should not be based on small samples.
- c. More items are generated for behavioral scales when the number of dimensions is increased. For example, there is the potential for nine dimensions to have up to 90 items or more.
- d. Raters appear to prefer a BARS format over a MSS format. It would probably not be useful to construct a MSS unless halo and leniency errors were anticipated.

4. Recommendations Regarding Use

- a. Scale development procedures will be strengthened if rater participation is included for BARS as well as other behavioral scale formats.
- b. BARS development procedures have resulted in a disproportionate rejection of mid-range items. Simple item intercorrelation procedures for the U<sub>s</sub> (universe score procedure) would increase the number of mid-range items. (DeCotiis, T. A. (1978). A critique and suggested revision of behaviorally anchored rating scales developmental procedures. Educational and Psychological Measurement, 38, 681-690.)
- c. Rigor in the developmental procedures for constructing various types of behavioral scales will influence and increase the reliability and validity of the scales more than the format.
- d. There appears to be a tendency to confound Thurstone scaling procedures with Likert scaling procedures which diminishes levels of reliability and validity for Thurstone scales. Researchers need to be aware of the differences between Thurstone and Likert scale development procedures when they are constructing BARS, BES, and BOS behavioral scales.
- e. To increase the MSS format acceptance by raters for the scoring system and item dimensionality, a coding system with face validity may be useful as well as training for the raters to explain the MSS rationale, and the procedures for carrying out the appraisal.
- f. MSS requires statistical analysis to ensure unidimensionality of the scales.

## F. Ranking Items

### 1. Definition and Examples

Ranking items call for the respondent to indicate the relative ordering of the members of a presented group of objects on some presumably discriminable dimension, such as effectiveness, saltiness, overall merit, etc. By definition, one does not have a scale by which the amount of difference between successive members is measured, nor is it implied in rank ordering that successive differences are even approximately equal. If respondents were being asked to give judgments on the size of intervals, the item would be something more than a ranking item.

Multiple choice items are so frequently used that one may inadvertently use this format when the ranking item format would provide more complete and reliable information. Item 1 in Figure IV-C-1 illustrates this point. Since a preponderance of respondents would check "protection" as a helmet's most important characteristic, only a small remainder of responses would be available as a basis for ordering the other characteristics. Some of the other characteristics might be achievable without sacrificing protection, so it would be desirable to have a reliable ordering of their importance.

As the number of objects to be ranked increases, the difficulty of assigning a different rank to each object increases even faster. This means that reliability (repeatability) is reduced. To counter this, one may explicitly permit respondents to assign tied rankings to objects when the number of objects exceeds, say, 10 or more.

Examples of ranking items are shown in Figure IV-F-1.

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). These systems involve establishing a rank order scale that is converted to an interval scale. Procedures and recommendations for constructing rank ordering embedded in subjective workload assessment methods can be found in a number of sources including:

- a. 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.
- b. 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.

- c. Eggemeier, F. T., McGhee, J. Z., & Reid, G. B. (1983, May). The effects of variations in task loading on subjective work-load rating scales. Proceedings of the IEEE 1983 National Aerospace and Electronics Conference, Dayton, OH, 1099-1105.

Figure IV-F-1

Examples of Ranking Items

1. Rank the following three methods of issuing starlight scopes to an infantry squad. Assign a "1" to the most effective, a "2" to the second most effective, etc. Do not assign tied rankings.

Ranking	Basis of Issue
_____	Scopes issued to AMG and SL
_____	Scopes issued to AMG, SL, and one rifleman
_____	Scopes issued to all squad members

2. How important are each of the following factors to you? Assign a "1" to the most important, "2" to the second most important, etc. Assign a different number to each of the four factors.

_____	Type of furniture in the barracks
_____	Army pay
_____	Medical service to soldiers
_____	Choice of duty station

2. Advantages of Ranking Items

- The idea of ranking is familiar to respondents.
- Ranking takes less time to administer, score, and code than paired-comparison items do, and there is some evidence that the results of the two are highly similar.
- Ranking and rating techniques are generally comparable in terms of reliability.

3. Disadvantages of Ranking Items

- Ranking items such as item 1 in Figure IV-F-1 do not reveal the respondent's judgment as to whether any of the objects are effective or ineffective in an absolute rather than just a relative sense. To learn this, another question must be asked.

- b. Rank order scales originate from ordinal scale measurement. The categories in a rank order scale do not indicate how much distance there is between each category. Unequal distances are assumed. Rank order items do not permit respondents to state the relative amounts of differences between alternatives.
- c. The results from ranking items are open to question if the basis for ranking was not clear to the respondents.
- d. Ranking is generally less precise than rating.

4. Recommendations Regarding Use

- a. Rank order scales are appropriate for analyzing data that meets the requirements of ordinal measurement scales.
- b. There are some situations where the intent of the questionnaire developer is best served with the use of one or more ranking items. Generally, however, rating scale items are probably preferable.
- c. Rank order scales and rating scales are more cost effective and time effective to use than paired-comparisons.
- d. Individuals tend to more frequently use one end of a list than the other end while ranking. To counteract this bias, it is possible to develop two or more versions of the list by randomly ordering the lists.
- e. It is possible to combine rank ordering with other methods, such as task analysis, to isolate critical components of a job. This information can be transformed into a performance measurement system, or can be used to modify military training.
- f. Analysis of the data for test-retest reliability performed on rank order, paired-comparison, and Likert scales varied depending on whether a Spearman rho or Kendall's tau was used. Kendall's tau may be a more appropriate measure of reliability for rank order measures.

G. Forced Choice Items

1. Definition and Examples

It would appear that any multiple choice item could also be called a "forced choice" item because, after all, the respondent is expected to choose one of the response alternatives. The instructions and/or the presence of an administrator put some degree of social pressure - social force - on the respondent. However, if a multiple choice item includes an "I don't know" response alternative, the pressure/force is almost totally removed. Likewise, on a rating scale item, the inclusion of a "neutral" or "borderline" response category allows the respondents to answer without committing themselves.

So, for some questionnaire developers - in particular those who produce "forced choice self inventories" (see references) - a "forced choice" item strictly refers to one where the respondents must commit themselves. They may have to select one of a pair of choices, or two of three, or two of four. These three cases are illustrated in Figure IV-G-1.

2. Advantages of Forced Choice Items

- a. Studies have indicated that reliabilities and validities obtained from the use of forced choice items compare favorably with other methods.
- b. The forced choice method has been used by a number of investigators in an attempt to control the tendency of individuals to answer self-report inventories in terms of response sets rather than giving "true" responses. (Response sets are discussed in Chapter XII.)

3. Disadvantages of Forced Choice Items

- a. Respondents sometimes balk at picking unfavorable statements, or at being forced to make a choice.
- b. Forced choice items take more time to develop than some other types of items.
- c. Paired-comparison items, where all phrases are paired, take more time to administer, score, and code than do ranking items. Results from the two, however, may have a linear relationship.

Figure IV-G-1

Examples of Forced Choice Items

1. Check one of the following two statements that is more characteristic of what you like.  
 I like to travel.  
 I like to meet new people.
2. Check one of the two following statements that is more characteristic of yourself.  
 I am honest.  
 I am intelligent.
3. Look at the following three activities. Mark an "M" by the one you like the most, and an "L" by the one you like the least.  
 Play baseball  
 Go to the craft shops  
 Attend boxing or wrestling matches
4. From the following four statements, check the two that are most descriptive of your unit commander.  
 Serious-minded  
 Energetic  
 Very helpful  
 Gets along well with others

- d. There is some question as to whether forced choice items overcome the biases or errors they are supposed to correct.
- e. Some investigators have concluded that the generalization that self-report forced choice inventories are more valid than single stimulus forms of the same tests is not supported by a critical consideration of the relevant evidence.

Procedures for constructing forced choice items, and evaluative comments about them, can be found in a number of sources including the following:

- a. Guilford, J. P. (1954). Psychometric methods (2nd ed.). New York: McGraw-Hill.
- b. Nunally, J. C. (1967). Psychometric Theory. New York: McGraw-Hill, pp 484-485.
- c. Sisson, E. D. (1948). Forced choice--the new Army rating. Personnel Psychology, 1, 365-381.

4. Recommendations Regarding Use

When test participants are deliberately given relevant experience with the operation of a weapons system, vehicle, or other system, the "I don't know" response alternative should normally be deleted from items that seek the participants' evaluations of the system.

## H. Card Sorting Items/Tasks

### 1. Definition

With card sorting items/tasks, the respondents are given a large number of statements (e.g., 75), each on a slip of paper or card. They are asked to sort them into, say, nine or eleven piles. The piles are in rank order from "most favorable" to "least favorable" or "most descriptive" to "least descriptive," etc., depending upon the dimension to be used. Each pile usually is to have a specified number of statements placed into it as required to form a rough normal distribution. However, some investigators have argued that forcing a given distribution is not necessary. Ordinarily each pile is given a score value which is then assigned to the statements placed into it.

An extensive discussion of the use of card sorts (or, more generally, Q-technique and its methodology) appears in: Stephenson, W. The study of behavior. Chicago: University of Chicago Press, 1953.

### 2. Advantages of Card Sorting Items/Tasks

- a. Card sorts appear to be capable of counteracting at least some of the biasing effects of response sets. (Response sets are discussed in Chapter XII.)
- b. Some investigators believe that card sorting is a fast and interesting method of obtaining valid and reliable interview data.
- c. With card sorts, the respondents can shift items back and forth if they wish to do so.
- d. The card sort has greatest value when a comprehensive description by a single individual is desired.
- e. Card sorts also have value for obtaining complex descriptions which can be compared systematically.
- f. They can be used to obtain rating information on any issue.

### 3. Disadvantages of Card Sorting Items/Tasks

- a. Card sorting items/tasks may take more time to construct than other types of items, and they generally take more time to administer and score.

b. Card sorts are more involved to administer than other types of questionnaire items.

4. Recommendations Regarding Use

Some authors think that card sorting is the method of choice if testing time is available. Its greatest value seems to be its ability to provide a comprehensive description by a single individual, or to obtain complex descriptions which can be systematically compared. Since it is more awkward to administer and score than other types of items, its use in Army field test evaluations is limited.

## I. Semantic Differential Items

### 1. Definition and Examples

The semantic differential technique was initially developed as a general method of measuring meaning, and with it the meaning of a particular concept to a particular individual can be specified quantitatively. The technique has also been used to measure attitudes and values, particularly in the marketing area. In using the technique, the respondent is presented with a number of bipolar rating scales, usually but not always having seven points. The two ends of each scale are defined by adjectives. The respondent is given a set of such scales, and is asked to rate each of a number of objects or concepts on every scale. To aid in interpretation, some scale coding can be used, usually numbers in a direct numerical sequence such as 1 through 7. Other more extensive scoring can be used, and results can be factor analyzed to search for the basic dimensions of meaning. However, the usefulness of the semantic differential as a research tool stems from the ability of the procedure to probe into both the content and the relative intensity of respondents' attitudes.

Examples of semantic differential items are given in Figure IV-I-1. A recommended text on the semantic differential is Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). The measurement of meaning. Urbana, Ill., University of Illinois Press. Norms have been collected on 20 scales for 360 words. They are reported in Jenkins, J. J., Russell, W. A., & Suci, J. (1958). An atlas of semantic profiles for 360 words. American Journal of Psychology, 71, 688-699.

### 2. Advantages of Semantic Differential Items

- a. Evidence on the validity, reliability, and sensitivity of the scales has been offered.
- b. Using some adjectives that do not seem appropriate to the concept under investigation may uncover aspects that reflect an attitude or feeling tone even though the respondent cannot put it into words.
- c. Semantic differential items can be used to study the relative similarity of different concepts to the respondent, and to study changes over time.
- d. Semantic differential items are relatively easy to construct, administer, and score.

Figure IV-I-1

Examples of Semantic Differential Items

1. Place an X in each of the following rows to describe your assessment of the M16 rifle.

Reliable \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_ Unreliable  
Heavy \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_ Light  
Good \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_ Bad  
Slow \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_ Fast  
Adequate \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_ Inadequate

2. Place an X in each of the following rows to describe your assessment of the ABC helmet.

Reliable \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_ Unreliable  
Heavy \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_ Light  
Good \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_ Bad  
Slow \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_ Fast  
Adequate \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_: \_\_\_ Inadequate

3. Disadvantages of Semantic Differential Items

- a. If care is not taken, the two adjectives chosen for the extremes will not define some kind of scale or dimension between them.
- b. The value of semantic differential items depends on the suitable choice of the bipolar adjectives and concepts.
- c. There is a potential response error present in the respondents' interpretations of the meaning of the end-point descriptions. However, there appears to be a balancing out over a number of administrations.
- d. There is the possibility of a socially desirable response set when personality traits are measured with the semantic differential.

4. Recommendations Regarding Use

- a. There are a number of investigators that advocate the use of the semantic differential. Others, however, have questioned whether it may be a rather complicated way of developing a measure that is more readily and reliably secured by other means. It is reasonable to assume that the technique could easily be expanded to identify attitudes and the intensity of the attitudes toward the attractiveness of a particular military specialty, the capacities of a specific piece of equipment to perform, or any other characteristic set which can be described by bipolar adjectives. However, since the analysis of sets of semantic differential items is somewhat involved, the technique has not been widely used for routine Army field test evaluations.
- b. Semantic space for the concepts of evaluation, potency, and activity are fairly stable across studies, and have maintained reliability over time. Because of the stability of the scale, it is possible to vary instrument format as well as rating instructions and maintain the viability of the scale. To ensure the soundness of the scale, developmental procedures need to include testing the instrument in the context area for which it was designed.
- c. In the early stages of development for the semantic differential, it is possible to identify potential bipolar anchors using Roget's Thesaurus as a source in addition to the subjects' concepts of terms that have semantic stability. Initial pools of items can be reduced through judgment agreement, factor analysis, and cluster analysis.
- d. Semantic differential scales can be anchored with phrases, adjectives, or adverbs.
- e. The number of scale points used with the semantic differential can vary, and still retain the integrity of the instrument. An acceptable range in the scale would be between five and twelve points. Each completed survey would have all items with the same number of scale points. For example, two questionnaires could be designed, one with seven scale categories and the other with nine scale categories.
- f. Social desirability response sets can be controlled by careful construction of the bipolar scales. Adjectives can be selected that reflect a common trait to control the influence of social desirability.

J. Other Types of Items

1. Checklists

Checklists are instruments in which responses are made by checking the appropriate statement or statements in a list of statements. Examples are shown in Figure IV-J-1.

Figure IV-J-1

Examples of Checklists

1. Which of the following are important to consider when deciding whether or not to make a career of the Army? Check all that apply.

- Leadership of NCOs
- Opportunity for promotion
- Playboy magazines in the Post Exchange
- Latrine in crafts shops
- Army pay
- Choice of duty stations
- Civilian opinion of Army
- Reenlistment bonuses
- Hours of work in a work week

2. Please check all the characteristics which Backpack A possesses.

- Durability
- Lightness
- Wearing comfort
- Accessibility of items
- Ease of putting on and taking off
- Other (specify): \_\_\_\_\_

Checklists can be used in conjunction with interviews to serve as a cue to the interviewer. Administration of a checklist combined with an interview of critical areas identified on the checklist could reduce interviewing time. Examples are shown in Figure IV-J-2.

Figure IV-J-2

Example of Checklist Pertaining to  
 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 check Not Applicable (NA).

<u>Equipment</u>	<u>Yes</u>	<u>No</u>	<u>NA</u>	<u>Comment</u>
1. Crane	_____	_____	_____	_____
2. Crane remote controls	_____	_____	_____	_____
3. Crane onboard controls	_____	_____	_____	_____
4. Winch	_____	_____	_____	_____
5. Winch controls	_____	_____	_____	_____

This checklist/interview could serve as the foundation for generating other, more refined instruments. The checklist/interview is another way of eliciting information from a subject matter expert group.

Compared to rating scales, which give a numerical value to some sort of judgment, checklists are relatively crude. They are, however, quite useful when scaled information is not needed. Checklists also are useful when information is needed to determine which of several issues are significant to a respondent. Other issues regarding the use of checklists are as follows:

- a. Checklists should use terms like the respondent uses.
- b. Response set can be somewhat controlled if the respondent is asked to check a stated number of items, or if upper or lower limits are set.
- c. There is some evidence that a higher rate of claim or assertion is obtained from checklists than from open-ended items.

- d. It is usually not known if checklists cover the appropriate attributes.
- e. Adjective checklists are sometimes used, especially to elicit stereotypes about people or nations. They are similar to rating scales.

## 2. Matching Items

With matching items, the respondent is given two columns of items, and is asked to pair each item in the first column with an associated item in the second. In general, it is not desirable to have the same number of items in each column. Both sets of items should constitute a homogeneous set, and any item in the second column should look like it could go with any item in the first column.

Matching items are best used in achievement testing. Since they have little utility in Army field test evaluations, they are not discussed in greater detail in this manual.

## 3. Arrangement Items

With an arrangement item, a number of statements are presented in random order, and the respondent arranges them in a new order according to his/her judgment and the guidance received. For example, steps in a sequence of events or procedures may be rearranged in order of occurrence or performance. Or, causes may be rearranged in order of importance in bringing about a certain effect.

There may be some situations where arrangement items may be useful in Army field test evaluations; however, the scoring of the items is difficult. The use of such items is, therefore, extremely limited.

## 4. Formats Providing for Supplementary Responses

The questionnaire writer is not limited to the major item formats described in this chapter. Formats providing for supplementary responses can also be used. Examples are shown in Figure IV-J-3.

Figure IV-J-3

Examples of Formats Providing for Supplementary Responses

1. The starlight scope is able to detect aggressor movements:

- very effectively.
- effectively.
- borderline.
- ineffectively.
- very ineffectively.

Explain: \_\_\_\_\_  
\_\_\_\_\_

2. What style of leadership was used by the most effective squad leader you served under? (Check one)

- democratic and friendly
  - friendly with most; authoritarian with the others
  - sometimes authoritarian; sometimes acts like one of the men
  - usually authoritarian; avoided making close friends
  - other (please describe) \_\_\_\_\_
- \_\_\_\_\_

Notice that the "other" response alternative in Example 2 allows the respondent in effect to make an open-ended item out of a multiple choice item. Few test respondents, however, elect to do this. Inclusion of the supplementary or write-in option commits you to extra data reduction and analysis effort that would have been unnecessary had you anticipated and included all reasonable response alternatives.

## Chapter V: Attitude Scales and Scaling Techniques

### A. Overview

At times, the questionnaire developers will wish to treat the total group of items on a questionnaire as a single measuring scale, and from them obtain a single overall score on whatever they are interested in measuring. This is a common practice, especially with the measurement of attitudes. A typical attitude scale is composed of a number of questions/statements selected and put together from a much larger number of questions/statements according to certain statistical procedures. Some of these procedures, called scaling techniques, are discussed in this chapter.

A distinction is needed, however, between two ways in which the term scale is used in this manual. An attitude scale could be constituted of items each one of which employs a response scale. Aspects of response scales are discussed in Chapter VII on "Response Anchoring." A component of score could be achieved on each item. Adding these item scores together - which means considering the whole set of items as a scale - produces a total attitude score for the individual respondent.

There are, generally speaking, two general methods for the construction of scales such as attitude scales. The first method makes use of a judging group and one of the psychological scaling methods developed by Thurstone, as discussed in Section V-B. It results in a set of statements being assigned scale values on a psychological continuum. The continuum may be favorableness-unfavorableness, like-dislike, or any other judgment. The psychological scaling methods, therefore, have considerably greater application than for the scaling of attitudes. They can be used to scale statements or objects. They have been used, for example, to determine the perceived favorableness of words and phrases commonly used as rating scale response alternatives, as discussed in Chapter VIII.

The second general method is based on the direct responses of agreement or disagreement with attitude statements and does not result in a set of statements being assigned scale values on a psychological continuum. Both the Likert and Guttman scales discussed in Sections V-C and V-D are examples of this latter method.

For information (relating to attitude scaling and scaling techniques) beyond that contained in this manual, the following references may be consulted.

1. Babbitt, B. A., & Nystrom, C. O. (1985). Training and human factors research on military systems. Questionnaires: Literature survey and bibliography. Fort Hood, TX: Army Research Institute for the Behavioral and Social Sciences.

2. 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.
3. Edwards, A. L. (1957). Techniques of attitude scale construction. New York: Appleton-Century-Crofts.
4. 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.
5. 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.
6. 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.
7. Guilford, J. P. (1954). Psychometric methods (2nd ed.). New York: McGraw-Hill.
8. Gulliksen, H., & Messick, S. (Eds.) (1969). Psychological scaling: Theory and applications. New York: John Wiley.
9. Lemon, N. (1974). Attitudes and their measurement. New York: John Wiley.
10. 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.
11. 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.
12. Nunnally, J. C. (1967). Psychometric theory. New York: McGraw-Hill.
13. Thurstone, L. L. (1959). The measurement of values. Chicago: University of Chicago Press.
14. Torgerson, W. S. (1958). Theory and methods of scaling. New York: John Wiley.

## B. Thurstone Scales

This section discusses three scaling methods developed by L. L. Thurstone. Thurstone investigated rank order scales and how to compare psychological variables. He developed the law of comparative judgment with an underlying assumption that 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. Thurstone generated three new scaling methods based on his law of comparative judgment. The three scaling methods are known as equal appearing intervals, paired-comparison, and successive intervals. For additional detail, see the texts referred to in Section V-A.

### 1. Method of Equal Appearing Intervals

Thurstone's method of equal appearing intervals assumes that a group of statements of opinion about a particular issue could be ordered on a continuum of favorableness-unfavorableness, and that the ordering could be such that there appears to be an equal distance between the adjacent statements on the continuum.

The following steps are followed in the method of equal appearing intervals:

- a. From the literature or pilot interviews, a large number of statements (100 to 200) are compiled about the attribute or object of an attitude under study. Irrelevant, ambiguous, or poorly worded statements would not be selected.
- b. A number of judges, at least 50, are obtained. They should be similar to those individuals who will respond to the final statements on the questionnaire. The judges independently sort each statement into one of 11 piles. The first pile is defined as "Unfavorable" or "Most unfavorable," the middle or sixth pile is defined as "Neutral," and the eleventh pile is defined as "Favorable" or "Most favorable." The other piles are left undefined. The judges are told that the intervals between piles or categories are to be regarded as subjectively equal. They are also instructed to ignore their own agreement or disagreement with each item, and to judge each item in terms of its degree of favorableness-unfavorableness.
- c. The scale value for each item is usually determined by computing its mean or median, over all judges.
- d. Twenty to 25 statements with little dispersion in their scale values are then selected for use. The statements are selected so that the intervals between statements' scale values are approximately equal and/or are relatively equally spaced on the psychological continuum.

- e. The finally selected statements are usually placed in random order for presentation to respondents. The respondents are asked to indicate which statements they agree with, and which they disagree with.
- f. The respondent's score is the mean or median scale value of those statements for which he/she marked "Agree."

Some considerations for use of the Equal Appearing Intervals method are:

- a. The method of equal appearing intervals is designed to provide an interval scale as its output. The scale is at least ordinal (ranked).
- b. The method is useful when there are a large number of statements involved.
- c. Scale values from widely differing groups of judges appear to correlate highly with one another so long as judges with extreme views are eliminated.
- d. Graphic or numerical rating scales can be used by the judges instead of having the statements sorted into piles. Though 11 categories are usually used, some other number can be employed.
- e. There have been some psychometric questions about the unidimensionality of Thurstone scales. Even though research has been mixed as to which scaling methods are best, there is some evidence that Likert and Guttman scales may be sounder. Actual scale format does not seem to be as important as the actual developmental procedures in the construction of the scale.

## 2. The Method of Paired Comparisons

Thurstone developed a procedure for deriving an interval scale based upon what has been called the Law of Comparative Judgment. Basically, it is a method by which statements such as "A is stronger than B," "B is stronger than C," etc., are used to provide a scale with interval properties. The objects or statements to be ranked are presented two at a time, and the respondent is asked to choose between them. All possible combinations of pairs have to be presented. Hence the procedure becomes very cumbersome when there are more than 15 or so items. The determination of scale values is also laborious. Since the procedure is not used much in applied research, additional detail is not presented here.

## 3. The Method of Successive Intervals

The method of successive intervals is similar to the method of equal appearing intervals. However, no assumption is made concerning the psychological equality of the category intervals.

It is only assumed that the categories are in correct rank order and that their boundary lines are relatively stable. The procedure involves estimating the widths of the categories along the psychological continuum. From these reference points, the scale values of the statements can be obtained. Research has shown that there is a linear relationship between scales constructed by the method of paired-comparisons and by the method of successive intervals.

#### 4. New Applications for Thurstone Scales

When Thurstone developed the law of comparative judgment, his scaling techniques were considered a major advancement. Thurstone 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 into 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). These systems involve establishing a rank order scale that is converted to an interval scale. More research will be required to determine how functional, reliable, and valid these new procedures will be. The procedures for embedding rank order methods into other scales is complicated and beyond the scope of this manual.

### C. Likert Scales

The Likert method of scale construction was developed because the Thurstone procedures require extensive work and make assumptions regarding the independence of item statements. The Likert method assumes that all statements reflect the same attitude dimension and are hence related to each other. The Likert approach does not assume equal intervals between the scale values. It is sometimes called the method of summated ratings.

The steps in Likert scale construction are as follows:

#### 1. Item Construction

Design an initial set of items to measure an attribute. Statements are classified in advance as "Favorable" or "Unfavorable." No attempt is made to find an equal distribution of statements over the whole range of the attitude of concern, and no attempt is made to scale the statements.

#### 2. Item Selection

Likert proposed the use of correlation analyses and analyses based on the criterion of internal consistency to evaluate the ability of individual items to measure an attribute.

- a. A pretest is conducted. In the pretest, the respondents indicate their degree of agreement with every statement, usually using five response alternatives: strongly agree, agree, undecided, disagree, and strongly disagree. Each descriptor is assigned a numerical weight (e.g., 4, 3, 2, 1, 0) usually based on a given series of integers in arithmetical sequence. Each respondent is assigned a score that represents the summation of weights associated with each item checked.
- b. Criterion of internal consistency compares the difference between mean responses to an individual item compared to high and low subgroups. Subgroups consist of 25% of the respondents at each extreme of the scale.
- c. The criterion of internal consistency includes differences in subgroup size and different distributions of responses between subgroups.

- d. The t test provides an accurate indication of the degree to which an item differentiates between high and low subgroups.

$$t = (\bar{X}_H - \bar{X}_L) / \sqrt{(S_H^2/n_H) + (S_L^2/n_L)}$$

$\bar{X}$  = mean item response of subgroup

$S^2$  = item variance of subgroup

$n$  = size of subgroup

- e. The criterion of internal consistency analysis and the correlation analysis may lead to different conclusions regarding the selection of items. It is recommended that both types of item analyses be used to assist in determining which items to retain.
- f. Correlational analysis focuses on how strongly the item is related to the total scale score.

$$r_{i(T-i)} = (r_{iT} \sigma_T - \sigma_i) / \sqrt{(\sigma_T^2 + \sigma_i^2) - 2r_{iT} \sigma_T \sigma_i}$$

$r_{iT}$  = correlation between item and total score

$\sigma_T$  = standard deviation of the total score

$\sigma_i$  = standard deviation of the item score

The greater the number of items, the less each item will contribute to the variance of the scale. Each item will contribute more bias for scales that have only a few items.

- g. Each item is treated as a predictor of the respondent's total score. Items with low item-to-total correlations should be eliminated from the scale. Items that do not discriminate between groups with extreme attitudes (25% of the respondents at each extreme of the scale) should be eliminated. This procedure leaves us with the items that will comprise the final score.

### 3. Item Scoring

- a. Calculate scale scores by summing the response scores for each item given the following values. Favorable statements receive a value of 4 for "Strongly agree" and a value of 3 for "Agree." The midpoint response alternative "Undecided" receives a value of 2. Unfavorable statements receive a value of 1 for "Disagree" and a value of 0 for "Strongly disagree." High scores always indicate a favorable attitude, and low scores always indicate an unfavorable attitude.

- b. Interpretation of individual scoring is defined relative to the group. Each of the individual attitude scores is expressed as a deviation from the mean of the group. The score of any individual relative to the mean of the group is:

$$X - \bar{X}$$

$X$  = individual score

$\bar{X}$  = group mean

The scores are converted into  $Z$  scores by dividing each individual score by the standard deviation of the sample. A  $Z$  score will identify the position of the respondent's score in relation to the mean of the distribution. Using the curve as a distribution of observations, the  $Z$  score can describe the location of the score along the horizontal axis. A  $Z$  score distribution maintains the same shape as the set of raw scores from which it was derived.

$$z = \frac{X - \bar{X}}{S}$$

$Z$  scores indicate how many standard deviations the score lies above or below the mean. The mean is always zero, and the standard deviation of any set of  $Z$  scores is always 1.  $Z$  scores can be used to compare scores from different distributions so long as the distributions have approximately the same shape.

#### 4. Reliability of the Summated Scale

To compute the reliability of the Likert scale, the coefficient alpha is recommended.

$$\alpha = N\bar{r} / [1 + \bar{r}(N-1)]$$

$N$  = number of items

$\bar{r}$  = mean interitem correlation

The alpha coefficient provides an estimate of reliability based on the interitem correlation matrix.

Factors to be taken into consideration when deciding whether to use Likert scales include:

1. Likert scales take less time to construct than Thurstone scales. They are one of the most widely used scales for attitude surveys.
2. It is possible to construct scales by the Likert and Thurstone methods which will yield comparable scores.

3. Likert scales have only ordinal properties. If there is a large dispersion about a respondent's mean score, however, even those properties have limited meaning. If the sole purpose of a scaling procedure is to rank respondents according to the degree to which they hold some attitude, then Likert scales are efficient because of their ease of administration.
4. In addition to lacking metric properties, Likert summated scores lack a neutral point. The interpretation of a score cannot be made independently of the distribution of scores of some defined group. Only the summation of the items measure the attitude. Percentile or deviation-type norms can be calculated if the sample size is large enough.
5. For the same number of items, scores from Likert scales may be more reliable than scores from Thurstone scales.
6. Likert and Guttman scales both appear to be superior to Thurstone scales.

D. Guttman Scales

Guttman scaling was developed as an alternative to Thurstone and Likert methods of attitude scaling. Guttman's approach to scaling is known as scalogram or scale analysis. It is a deterministic model; it considers its scales are close to being rulers-measures of length. The essence of the method is to determine whether a series of statements can be appropriately scaled. An attempt is made to identify a set of statements which actually reflect a unidimensional scale and have a cumulative nature. When the goal is achieved, two or more persons receiving the same score will have responded in the same way to all of the statements.

As an example, the following four questions comprise a Guttman scale:

	Yes	No
a. The United Nations is mankind's savior	___	___
b. The United Nations is our best hope for peace	___	___
c. The United Nations is a constructive force in the world	___	___
d. We should continue our participation in the United Nations	___	___

The expected pattern of responses to these questions is "triangular."

<u>Item</u>	<u>Person</u>				<u>Scale Score</u>
	1	2	3	4	
a	x				1
b	x	x			2
c	x	x	x		3
d	x	x	x	x	4

This means that, for persons who answers yes to item "a," there is a high probability that they will answer yes to the other items. A person who says no to "a" but yes to "b" has a high probability of answering yes to the other items, and so on. The model anticipates that the perfect relationship between the scale score and the item score will be violated. The degree of deviation that is acceptable is established by criteria, and measured by a coefficient of reproducibility.

Guttman scaling is considered psychometrically more robust than Likert or Thurstone scaling. The coefficient of reproducibility (CR) could be used to evaluate the degree of scalability of empirical data. The Guttman model calls for assigning scale scores only when the coefficient of reproducibility (CR) is greater than .90. The formula is as follows:

$$\begin{aligned} \text{CR} &= 1.0 - (\# \text{ errors}) / \text{total responses} \\ &= 1.0 - (\# \text{ errors}) / [(\# \text{ items}) \times (\# \text{ respondents})] \end{aligned}$$

For example, a respondent who rates three items positively out of  $n$  items composing the Guttman scale would be considered to have responded to three specific items which would be considered the three items most acceptable to the population of respondents. The interpretation of a response to three items on a Likert scale would be that the respondent had rated favorably any three items of  $n$  stimuli.

The major steps in scalogram analysis are too complex to summarize here, but are found in some of the references in Section V-A. Procedures are available for:

1. Measuring the amount of error due to imperfect scalability.
2. Ordering the statements so that the response patterns provide the least amount of error.
3. Determining the extent to which the data approximate the perfect case.
4. Improving the scalability of the statements via category combinations, statement discarding, etc.

There have been many critics of scalogram analysis. Some feel that there is no really effective way of selecting good items by this approach. However, the procedure is considered useful if one is concerned with unidimensionality or if one wishes to examine small changes in attitudes. Guttman scaling is primarily used in the construction of attitude surveys as well as in the construction of mixed standard scales. It may be possible to construct other mixed standard scales for surveys that measure other factors in addition to job performance. It is laborious to construct Guttman scales. No instances of past use in field testing situations are known.

Even though Guttman's approach to scale analysis has not been used in field testing situations, it is being used by the armed services for other applications.

Adaptive testing is based on a Guttman method of scaling and adaptive testing is being investigated by the armed services. The Armed Services Vocational Aptitude Battery is being developed for computer-adaptive testing by the Navy Personnel Research and Development Center. 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.

Adaptive testing requires a large sample for its development. It has been primarily used as an ability test with multiple choice questions. There have been other types of applications such as for interviewing. 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 required for development.

E. Other Scaling Techniques

Numerous other scaling techniques and combinations of methods are reported in the literature. A discussion of them, however, is outside the current scope of this manual.



Chapter VI: Preparation of Questionnaire Items

A. Overview

Once a decision has been made regarding the type or types of items that are to be used in a questionnaire (see Chapter IV), attention must be given to the actual development of the items. This chapter addresses the following development topics: mode of questionnaire items; wording of items for both question stems and response alternatives; difficulty of items; length of question stem; order of question stem; number of response alternatives; and order of response alternatives. The related topic of response anchoring is considered in Chapter VII.

As used in this manual, a distinction has been made between a questionnaire item, a question stem, and response alternatives. A questionnaire item has both a question stem and response alternatives. The response alternatives are the answer choices for the question. (They are sometimes called "options.") The question stem is that part of the item that comes before the response alternatives.

B. Mode of Items

Questionnaire items are usually presented to a respondent in printed form. However, it is possible to present items or stimuli pictorially. There is some evidence that there are no significant differences in subjects' responses to verbal and pictorial formats. The evidence is conflicting, since anchoring endpoints with pictorial anchors for bipolar scales has proven difficult in establishing meaning. Researchers were not able to verify that the pictorial anchors were actually antonyms. This could affect the bipolar assumptions of the scales. Using a pictorial format may facilitate obtaining responses from respondents with limited verbal comprehension, who might have difficulty responding to questions employing lengthy definitions of concepts or objects. If pictures are used, they should be pretested for clarity of their presentation of the concept or object to be evaluated.

For group administration of a questionnaire with pictorial anchors, it would be possible to use color slides and rating forms with replicas of the slides. In cases where it is known that the respondents have very low reading ability, it may be desirable to present the questionnaire orally. A tape player-recorder may be used for this purpose also.

C. Wording of Items

The wording of questionnaire items is a critical consideration in obtaining valid, relevant, and reliable responses. Consider, for example, the following three questions that were administered by Payne (see reference below) to three matched groups of respondents:

- a. "Do you think anything should be done to make it easier for people to pay doctor or hospital bills?"
- b. "Do you think anything could be done to make it easier for people to pay doctor or hospital bills?"
- c. "Do you think anything might be done to make it easier for people to pay doctor or hospital bills?"

These questions differed only in the use of the words "should," "could," or "might," terms that are often used as synonyms even though they have different connotations. The percent of "Yes" replies to the questions were 82, 77, and 63, respectively. The difference of 19% between the extremes is probably enough to alter the conclusions of most studies.

A number of matters related to the wording of questionnaire items are considered in this section. Some of the suggestions made are based upon experimental research. Others are based upon experience, intuition, and commonsense. Several sources offering principles of question wording are:

- a. Roslow, S., & Blankenship, A. B. (1939). Phrasing the question in consumer research. Journal of Applied Psychology, 23, 612-622.
- b. Jenkins, J. G. (1941). Characteristics of the question as determinants of dependability. Journal of Consulting Psychology, 5, 164-169.
- c. Blankenship, A. B. (1942). Psychological difficulties in measuring consumer preferences. Journal of Marketing, 6, 66-75.
- d. Payne, S. L. (1963). The art of asking questions (Rev. ed.). Princeton, NJ: Princeton University Press.
- e. Schuman, H., & Presser, S. (1981). Questions and answers in attitude surveys: Experiments on question form, wording, and context. New York: Academic Press, Inc.

1. Formulation of the Question or Question Stem

a. General comments regarding items and question stems. Issues that should be noted concerning the general structure of questions and question stems are:

- (1) Question stems may be in the form of an incomplete statement, where the statement is completed by one of the response alternatives, or in the form of a complete question. See Figure VI-C-1 for examples.

Figure VI-C-1

Example of Question Form (Item 1) and  
Incomplete Statement Form (Item 2) of Stem

1. How qualified or unqualified for their jobs are most Army NCOs?  
(Check one.)
- Very well qualified
- Qualified
- Borderline
- Unqualified
- Very unqualified
2. Check one of the following. Most Army NCOs are:
- Very well qualified for their jobs.
- Qualified for their jobs.
- Borderline.
- Unqualified for their jobs.
- Very unqualified for their jobs.

The choice between these two methods should depend on which of the two permits simpler and more direct wording for the item in question. Not all of the items in a questionnaire need to be in the same form.

- (2) All questionnaire items should be grammatically correct.
- (3) All stems should be as neutrally expressed as possible, and the respondents should be permitted to indicate/select the direction of their preference. If this is not done, the stems may influence the response distribution. If the stems cannot be expressed neutrally, then alternate forms of the questionnaire should be used.
- (4) Respondents may not answer an item if they are not able to give the information requested. Therefore, care should be exercised in the wording of the question, so that it does not call for information not possessed by the respondents. If the respondent is not able to answer the item, the response option should permit the respondent to say he/she "doesn't know." The questionnaire designer should have determined during pretesting whether a "Don't Know" response option should be included.

b. Accuracy and completeness of question stems.

- (1) The stem of an item should be accurate, even though inaccuracies may not influence the selection of the response alternative.
- (2) The question stem, in conjunction with each response alternative, should present the question as fully as necessary to allow the respondent to answer. It should not be necessary for the respondent to infer essential points. An example of an insufficiently informative question stem is given as item 1 in Figure VI-C-2. It is insufficient in that no specification is given as to who should carry the scopes. (The response alternatives are also insufficient since the respondent is not allowed to say "None.") Two or three questions might be needed to obtain all the information desired. Item 2 in Figure VI-C-2 is one revision that makes the question stem sufficient.
- (3) Generally, materials which are common to all response alternatives should be contained in the stem, if this can be done without the need for awkward wording.
- (4) In forming questions which depend on respondents' memory or recall capabilities, the time period a question covers must be carefully defined. The "when" should be specifically provided.

Figure VI-C-2

An Insufficiently Detailed Question Stem, Plus Revision

1. How many starlight scopes should be issued to a rifle squad?

- 1
- 2
- 3
- 4
- 5

2. Place a check in front of each squad member's "name" below that you believe should be issued a starlight scope:

- |   |   |
|---|---|
| <input type="checkbox"/> Squad Leader       | <input type="checkbox"/> Fire Team 2 Leader |
| <input type="checkbox"/> Fire Team 1 Leader | <input type="checkbox"/> Automatic Rifleman |
| <input type="checkbox"/> Automatic Rifleman | <input type="checkbox"/> Grenadier          |
| <input type="checkbox"/> Grenadier          | <input type="checkbox"/> Rifleman           |
| <input type="checkbox"/> Rifleman           | <input type="checkbox"/> Rifleman           |

(5) Question stems and response alternatives should be worded so that it is clear what the respondent meant. Consider the question "Should this cap be adopted, or its alternate?" If the respondent answers "Yes," it would still be unclear which cap ("this cap" or "its alternate") should be adopted.

c. Positive versus negative wording.

- (1) Alternative wording can produce demonstrable effects on survey results.
- (2) There may be a tendency for the direction of the question stem to be chosen in the response alternative.
- (3) Studies have indicated that it is usually undesirable to include negatives in question stems (unless an alternate form with positives is also used for half of the respondents).