(4) Questions worded in positive terms are preferred by respondents to questions in negative terms (if alternate forms are not being used). Questions worded negatively may be confusing, or negative words may be overlooked.

(5) If it seems necessary to have a particular question in negative form, the negative word (e.g., not, never) should be underlined or italicized. Care should also be taken that there are no double negatives, as they are frequently misinterpreted.

(6) A question worded in negative terms can often be improved by rephrasing it in positive terms.

(7) There is evidence to indicate that positively-worded items may in some instances receive higher mean responses than negatively-worded items. However, these findings were not statistically significant. There are conflicting research results where positive and negative wording of items did not affect the responses.

d. Definite versus indefinite article wording. The indefinite articles, "a" or "an," would be used in a question such as "Did you see a demonstration of the new night vision device?" A comparable question using the definite article "the" would be, "Did you see the demonstration of the new night vision device?" There is some evidence that changing from "a" to "the" reduces the level of suggestibility of an item. However, there is not enough evidence to warrant a firm conclusion.

e. First, second, and third person wording. An example of a statement written in the first person is, "Army NCOs are understanding of my needs and problems." A statement in the second person is, "Army NCOs are understanding of your needs and problems," while one in the third person is, "Army NCOs are understanding of the needs and problems of their men." It is preferable that the framework of questions be consistent for all questions in a questionnaire, so that responses are comparable. A respondent's opinion of events affecting his/her own person is often quite different than his/her opinions of the effects of the same events on others. Hence, questions written in the first or second person may elicit entirely different responses than the "same" question written in the third person.

There are occasions where each person (first, second, or third) is appropriate. For example, the third person should probably be used when it is desired to elicit information that might be considered too personal for a person to answer about himself/herself. The third person may also be used in attempts to elicit information about the feelings inherent in a minority of respondents, but about which many more respondents may be

aware, such as in the statement, "The Army is ahead of most areas of civilian life in reducing racial discrimination." In other cases, the first or second person form is not applicable, such as in "The Army is essential for the defense of the country." Also, the use of the third person permits a far larger number of personnel to answer the questions, since some first person questions that are inapplicable to many individuals become applicable when in the third person. Instances may occur where respondents are asked a question twice, once to discover how they personally feel about the issue (using first or second person), and then to discover what they judge others' feelings on that issue are (using the third person). Some personally-worded items may be perceived as more specific to the experience of the respondent. This may possibly provide results that have greater accuracy for items that are non-threatening. Generally, however, the use of the third person appears preferable.

f. Loaded and leading questions. Loaded and leading questions should be avoided. Although the questionnaire writers may not deliberately attempt to distort the distribution of responses, they may sometimes do so unintentionally.

In Figure VI-C-3, item 1 should be revised to maintain neutrality by removing the adjectives applied to the rifles. It is true that the M16 weighs less and fires more rounds faster, but there are other characteristics (accuracy, lethality given a hit, etc.) that are not cited. Hence, the question is loaded because it only presents some of the data relevant to comparing the rifles.

Items 2 and 3 in Figure VI-C-3 show loading of a different type. In item 2, analysis of the available alternatives leaves the impression that the writer of the question thinks at least some should not have a full automatic selector. Analysis of the alternatives in item 3 leads to the suspicion that the writer of the question believes there should be at least one grenade launcher in the rifle squad, since a response alternative of zero grenade launchers was not provided.

There are many additional ways that questions can be loaded. One way is to provide the respondent with a reason for selecting one of the alternatives, as with the question, "Should we increase taxes in order to get better schools, or should we keep them about the same?" A question can also be loaded by referring to some prestigious individual or group, as in, "A group of experts has suggested...Do you approve of this, or do you disapprove?"

Figure VI-C-3

Examples of Loaded Questions

1. Which rifle do you prefer, the lighter, faster shooting M16 or the heavier, slower firing M14?

   _____ M16

   _____ M14

2. Should every rifleman in the rifle squad have a full automatic selector on his rifle?

   Yes_____

   No_____

   If no, how many should?_____

3. How many grenade launchers (M79) do you desire in the rifle squad?

   _____ 1

   _____ 2

   _____ 3

   _____ 4 or more

Leading questions are similar to loaded questions. Two examples are shown in Figure VI-C-4. The problem is that most people are reasonably cooperative and like to help. If they can figure out what is wanted, they will often try to comply. The items in Figure VI-C-4 were actually used in the collection of data in a field test. As might be expected, the impression received from an analysis of the results is that men are, in general, highly motivated, and use good noise discipline during movement. (These items also allow respondents to avoid criticizing, and to give socially desirable answers.)

Figure VI-C-4

Examples of Leading Questions

1. Do you think your men were pretty highly motivated on this exercise?

   Yes_____

   No_____

2. Were they pretty good at using good noise discipline during movement?

   Yes_____

   No_____

The best way to avoid loaded questions is to find a devil's advocate to review them or to pretest the items on someone who holds opposite or minority views. Another check is to ask yourself what you think, what someone who disagrees with you would think, and whether your response alternatives would give the respondents a chance to present their views.

Not every change in wording will have a significant effect on the item. This provides a measure of latitude in the design of the items. Blatant attempts to bias an item by tone of wording are not so likely to succeed. Research indicates that blatant language may have no effect on responses. There is no convincing evidence that respondents with strong attitudes toward a topic would be less influenced by the tone of wording than respondents who did not have a strong attitude toward the topic.

There are times when loaded questions probably should be used. This is when, without loading, the question would pose an ego-threat to the respondents, so that they might give an untruthful reply. The loading removes the ego-threat so that a more valid response can be obtained. An example might be, "Many people are not able to get as much schooling as they would like. What was the last grade you completed in school?"

g. Embarrassing or self-incriminating questions. Respondents should not be asked embarrassing or self-incriminating questions. Consider the question, "Did you clean your weapon regularly in Vietnam?" It is asking respondents who did not clean their rifles regularly to expose themselves to possible embarrassment. Thus, one would expect the percentage of "No" responses to fall short of the true percentage not cleaning their weapons "regularly."

Occasionally questionnaires cover topic areas that are sensitive, and may be perceived as threatening by the respondents. For this type of questionnaire, threatening questions elicit greater under-reporting when closed-end questions are used. Thus, open-ended questions are appropriate for threatening topics. Longer questions, using the language of the respondent, seems to decrease unwanted response effects for threatening questions. In addition, willingness to answer threatening questions is increased by assuring respondents that their answers will be treated confidentially. An example of a threatening question is presented in Figure VI-C-5 which illustrates a longer, open-ended item used with threatening content.

Figure VI-C-5

Example of a Threatening Question

Please describe and explain in your own words any problem in your unit that might be caused by the use of too much alcohol, marijuana, or hard drugs by upper-ranking officers, senior NCOs, or supervisors.

_____

_____

_____

_____

_____

h. Questions that ask respondents to go against basic inclinations.

Many people are reluctant to criticize, though they enjoy giving praise. Thus, a question that allows respondents to avoid criticism will bias their answers; similarly, a question that offers them the opportunity to criticize may bias responses because they will not wish to do so. Figure VI-C-6 illustrates this.

Figure VI-C-6

Example of a Question
Asking the Respondent to Criticize

1. Was your unit's use of fire and maneuver correct, and in accordance with current Army doctrine?

   Yes_____

   No_____

   If no, why not?_____

   _____

The question in Figure VI-C-6 asks the respondents either to criticize their unit or to avoid criticism. Some respondents might answer "No" if they have an important point to make. However, a substantial number of others will wash their hands of the whole affair and answer "Yes," although they might feel that performance was not completely correct.

i. Inclusion of different subjects into the same question. Compound questions should be avoided. These are questions that require a respondent to give the same assessment of two or more issues/characteristics or aspects of the subject. Respondents must be allowed to make separate assessments of each issue. Consider, for example, item 1 in Figure VI-C-7. Most respondents would probably want to rate completeness and accuracy differently, since in most situations research has shown that they are negatively correlated. Therefore, the two aspects of performance should be rated separately, as shown in items 2 and 3 of Figure VI-C-7.

Figure VI-C-7

Examples of Compound Questions and Alternatives

1. How complete and accurate was the surveillance information?

_____ Very satisfactory

_____ Satisfactory

_____ Borderline

_____ Unsatisfactory

_____ Very unsatisfactory

2. How complete or incomplete was the surveillance information?

_____ Very complete

_____ Fairly complete

_____ Borderline

_____ Fairly incomplete

_____ Very incomplete

3. How accurate or inaccurate was the surveillance information?

_____ Very accurate

_____ Fairly accurate

_____ Borderline

_____ Fairly inaccurate

_____ Very inaccurate

It may be noted that in item 2 of Figure VI-C-7 both "complete" and "incomplete" are included. Similarly, both "accurate" and "inaccurate" are in the stem of item 3. To use only one (e.g., "complete") in the stem would tend to inflate the number of respondents selecting that alternative.

j. Use of giveaway words. Avoid words which lead the careful thinker to respond in the negative, while others, thinking less carefully, respond in the positive. Consider for example the question, "Do you feel that your unit did its best in all contacts over the past six months?" One wonders if any unit can do its actual best, except very rarely. The word "all" makes this an even more difficult question to answer positively.

k. Ambiguous questions. Vague or ambiguous words or questions should be avoided. For example, the question "What is your income?" is not sufficiently specific. The respondents may give monthly or annual income, income before or after taxes, their income or the family income, etc.

As another example, consider item 1 in Figure VI-C-8.

Figure VI-C-8

Example of Ambiguous Question and Alternative

1. Did you clean your rifle regularly in Vietnam?

_____ Yes

_____ No

2. How often, on the average, did you clean your rifle in Vietnam?

_____ Every day _____ Once every three days

_____ Once every two days _____ Once every four days

_____ Other (please specify): _____

_____

Use of the word "regularly" without specification of the time interval between cleanings is a defect in the question. A respondent could justify a "yes" by thinking to himself/herself: "Sure, I cleaned it regularly - once every four months!" Because of the self-exposure involved, the questionnaire item approach to this topic is probably not capable of providing an accurate estimate, but rewording could still make the amount of underestimation less. So, if the data cannot be collected by field inspection, the revised questionnaire item could read like item 2 in Figure VI-C-8.

Items are sometimes loaded because the wording is ambiguous, coerces agreement, or uses jargon or technical words that are not understandable. Review of items for illogical response patterns may be useful when respondents have less education. Figure VI-C-9 illustrates items which were highly ambiguous in their wording. Some respondents did not consider the first item in a literal sense for its impact on subsequent items. This set of items obtained many illogical response patterns.

Figure VI-C-9

Example of Ambiguity of Wording

"Are there any situations you can imagine in which you would approve of a policeman striking an adult male citizen?"	YES, NO, NOT SURE
"Would you approve if the citizen . . ."	
A. "had said vulgar and obscene things to a policeman?"	YES, NO, NOT SURE
B. "was being questioned as a suspect in a murder case?"	YES, NO, NOT SURE
C. "was attempting to escape from custody?"	YES, NO, NOT SURE
D. "was attacking the policeman with his fists?"	YES, NO, NOT SURE

When items are long and negatively worded, they may create ambiguity. This ambiguity seems to result in an increased number of responses in the middle alternative. Pretesting the items would provide the opportunity for modification of items by obtaining feedback from the respondents on issues related to the complex meaning of any technical words, and any multiple meanings of words. Use the language of the respondents in developing and refining items.

2. Formulation of the Response Alternatives

When formulating the response alternatives portion of a questionnaire item, the following points should be kept in mind:

a. All response alternatives should follow the stem both grammatically and logically, and, if possible, be parallel in structure.

b. If it is not known whether or not all respondents have the background or experience necessary to answer an item (or if it is known that some do not), a "Don't know" response alternative should be included.

c. When preference questions are being asked (such as "Which do you prefer, the M16 or the M14 rifle?"), the "No preference" response alternative should usually be included. The identification of "No preference" responses permits computation of whether or not an actual majority of the total samples are pro or con.

d. Respondents with a low educational level have a propensity to use the "Don't know" response alternative.

e. When the "Don't know" response alternative is used, it should be set apart from other responses to avoid confusing it with the endpoint or the midpoint of the rating scale.

f. Content items can be developed which will indicate whether a subject has knowledge regarding the topic in question. If the subject has little topic knowledge, and there is not a "Don't know" category, there is the potential for greater rating error.

g. The use of the "None of the above" option or variants of it, such as "Not enough information," is sometimes useful.

h. The option "All of the above" may on rare occasions be useful. It seems more appropriate to academic test questions than to the questioning of field test participants.

i. For most items, the questionnaire writer desires the respondent to check only one response alternative. Use of the parenthetic "(Check one.)" should eliminate the selection of more than one alternative. It is very important to make it clear to the respondents that they may check more than one alternative in those fairly rare instances where the questionnaire writer does wish to permit this.

j. In some instances, response categories as long as a sentence may be more desirable than short descriptors. In rare cases, numbers may be used without verbal descriptors, if the numbers have been previously defined. It does not seem to matter if the response alternatives are numerical, verbal (one word), or phrases. No one type of response alternative has proven superior to another.

k. When the quality of the item is high and the data is available, response alternatives can be selected which have standard deviations less than 1.00 (see Section VIII-E).

l. There is some evidence that responses to scales labeled at only the extreme ends have been skewed toward the positive end of the scale. Fully labeled scale points may encourage a more balanced response distribution.

m. Number of response alternatives is discussed in Section VI-G, order of response alternatives in Section VI-H, response anchoring in Chapter VII, and the order of perceived favorableness of commonly used words and phrases in Chapter VIII.

3. Expressing Directionality and Intensity in Stem Versus Response Alternatives

In item 1 of Figure VI-C-10, directionality (in this case, satisfaction) is expressed in the question stem.

Figure VI-C-10

Alternate Ways of Expressing Directionality and Intensity

1. The M16 is a satisfactory rifle.

_____ Agree

_____ Disagree

2. The M16 is

_____ a satisfactory rifle.

_____ an unsatisfactory rifle.

3. The behavior of civilian employees of the PX toward enlisted personnel is extremely offensive.

_____ Agree

_____ Disagree

4. The behavior of civilian employees of the PX toward enlisted personnel is

_____ very offensive.

_____ somewhat offensive.

_____ neutral.

_____ somewhat pleasant.

_____ very pleasant.

In item 2, the directionality is expressed in the response alternatives. In item 3, the stem contains terms of intensity and directionality, while these terms are located in the response alternatives in item 4. Item 2 is preferred to item 1, and item 4 is strongly preferred to the item 3 approach. The rationale for this preference is similar to the discussion of positive versus negative terms. Those who check "Disagree" to item 3 have not been permitted to indicate what it is they would agree with, (e.g., those who feel employees are offensive but not extremely offensive would have to check "Disagree," as would those who feel employees are very pleasant), whereas the construction of item 4 does permit them to do so. It would take five versions of item 3 to correct this deficiency and achieve the coverage of opinion incorporated by the response alternatives of item 4.

D. Difficulty of Items

1. One of the major recommendations advanced by almost every general source on how to write sound questionnaires is "keep it simple." Logic dictates that words used in surveys should not have multiple meanings, nor should they be beyond the level of vocabulary of the typical respondent. Words, phrases, and sentence structures that the respondent can understand should be used.

   Consider item 1 in Figure VI-D-1. It contains too many hard to understand words. Many respondents would have difficulty understanding either the question or the response alternatives. In the revision in item 2, the words have been simplified and a "catch-all" open-ended response alternative added (to catch all other reasons).

Figure VI-D-1

Example of Hard to Understand Item and Alternative

1. In the highly specialized counterinsurgency environment represented by the basically internecine affair in Vietnam, what would you say should represent the basic essence of our rationale for continuation of our involvement?

   _____ Prolongation of attrition of enemy forces, in order to reduce the level of threat to South Vietnam.

   _____ Orderly transfer of military responsibility to the host country, in order to produce stabilized competency to deal with any future internal distrubances.

2. What is our main reason for staying in Vietnam? (Check one)

   _____ To reduce the threat to South Vietnam by continuing the destruction of enemy forces.

   _____ To assure South Vietnam's survival while it takes over responsibility for its own protection.

   _____ Other (specify) _____

   _____

It should not be assumed that the respondent will understand what the question writer is talking about. Consider, for example, the question "Which do you prefer, dichotomous or open questions?" The odds are that a fairly substantial number of people would not be able to define these two question types. However, if they are asked this question, they will be happy to choose. The point is that people will not volunteer their ignorance of something, although they may admit it if you ask them. However, this caution goes beyond ignorance of an issue. Another problem is that the specialists wording the question may simply have an unusual command of their own language. Scientific jargon has been criticized. Perhaps overlooked is the fact that there are other kinds of jargon, too. The question askers have a responsibility to make themselves understood. One way of screening for individuals who do not have a basis for providing the information needed is to include one or two pure information questions. Plan to discard questionnaire returns from respondents who cannot answer the information questions correctly. However, our usual policy should be to throw out or revise items that are not understandable, rather than to throw out the responses of the people who can't understand the item.

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

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

2. Ways of Measuring Item Difficulty

Various procedures exist for determining the difficulty or reading comprehension level of printed material. Such a discussion is, however, beyond the scope of this manual. Sources that may be consulted include:

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

b. Dale, E., & Chall, J. S. (1948). A formula for predicting readability. Educational Research Bulletin, 27, 11-20, 37-54.

c. Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32, 221-233.

d. Fry, E. (1968). A readability formula that saves time. Journal of Reading, 11, 513-516.

e. Lorge, I. (1944). Predicting readability. Teachers College Record, 45, 404-419.

f. Schaefer, B. A., Bavelas, J., & Bavelas, A. (1980). Using echo technique to construct student-generated faculty evaluation questionnaires. Teaching of Psychology, 7(2), 83-86.

g. Thorndike, E. L., & Lorge, R. (1944). The teacher's word book of 30,000 words. New York: Columbia University Press.

E. Length of Question/Stem

This section notes some considerations about the length of question stems. There is little research in this area to guide the questionnaire writer. See Section IX-C regarding questionnaire length.

1. It is sometimes desirable to break the question stem into two or more sentences when the sentence structure would otherwise be unnecessarily complex. For instance, one sentence can state the situation, and one can pose the question. Lengthy question stems that try to explain a complicated situation to the respondent should be avoided. If the respondents are not aware of the facts presented, they may become more confused or biased than enlightened, and their opinion would not mean much.

2. Longer open-ended questions do not necessarily produce a greater amount of and more accurate information than shorter ones. However, it may take more words to achieve a proper focus.

3. Questionnaire developers have a tendency to use long question stems with true-false questions when "True" is the correct answer. Respondents often detect and react to this tendency. Field test questionnaires, however, should make relatively little use of "True" and "False" response alternatives. These alternatives are more appropriately used when testing whether respondents have acquired a required proficiency level, for example, the ability to visually recognize a given type of enemy aircraft.

4. To obtain higher reporting levels by respondents when threatening questions are asked about their behavior, longer items may be best. Items with 30 or more words have achieved best results. Items with fewer words (less than 30) have not elicited reporting levels which were as high. One of the longer items had 49 words, and the content was about the use of drugs.

F. Order of Question Stems

There are two issues to consider regarding the order of question stems. The first has to do with the order of questions within a series of items that are designed to explore the same topic or subject matter or related subject matter areas. The second has to do with the order of different groups of questions when the groups deal with fairly separate topics or subject matter areas. For example, one group of questions may deal with factual items, while another may deal with attitudes. If items bearing on the same point are presented in succession, the respondent can proceed more readily through them. Thus, this is usually a desirable practice. An exception arises when one wishes to check the consistency of the respondents. To do this, two (or more) similar items are included, but at widely different points in the questionnaire.

1. Order of Questions Within a Series of Items

a. It is often recommended that the order of questions on an instrument be varied or assigned randomly to avoid one question contaminating another. The view is that the immediately preceding question or group of questions places the respondent in a "mental set" or frame of reference. For example, asking respondents a general question about their feelings regarding automobile exhaust pollution might influence responses to the question, "Do you prefer leaded or nonleaded gasoline?" Questionnaires are plagued by contextual effects attributed to item ordering. Respondents lacking in experience of the content area may change their responses as they progress through the questionnaire, since they may learn from previous items (order effects). This may damage the face validity of the responses to the initial items. Yet, the meaning of the items would be changed if they were separated from their topic areas. The current state-of-the-art for context effects suggests that all items which are interrelated by content area may be affected by context effects. There is currently no way to predict which items will have context effects.

b. Sometimes it is recommended that broad questions be asked before specific questions. The rationale for this approach is that the respondent can more easily and validly answer specific questions after having had a chance to consider the broader context. Also, asking the specific questions first could influence the response to the broader question. The quality of responses to questions on a questionnaire will be determined by the respondent's background and knowledge of the topic area. A series of specific questions (versus general questions) will provide information about whether the respondent understands the content of the questions. It should expose any logical inconsistencies in response patterns. Respondents with limited or no experience regarding the content area may deviate from the logical response pattern. Their answers to questions may change as they become more familiar with the topic through order effects. Early responses may not have face validity.

General and specific questions were empirically examined for order effects. The order of the questions did not appear to effect the way respondents marked the response alternatives. It is proposed that a stronger survey instrument may be provided by assigning general items first, followed by specific items on related topic areas. However, questions which are specific are preferred over general type questions. Contextual effects can be minimized by developing questions which are specific in content. Minimize the number of general questions.

c. The order of questions within a series of items will also depend upon whether filter questions are needed. A filter question is used to exclude respondents from a particular sequence of questions if those questions are irrelevant to them. For example, if a series of items were asked about different kinds of weapons, a "No" response to a question such as "Have you ever used the M14 rifle?" might be used to indicate that the respondent should skip the following question(s) about the M14.

When filter questions are used by an interviewer, they can reduce interviewing time. Clear branching instructions are imperative for the interviewer. Filter questions used to branch in mail surveys or group-administered questionnaires have the potential to increase non-response rate for questions which follow a branch. Items following a branch tend to receive a lower response rate. Response rate for individuals over 60 years of age are even lower. There are alternatives to branching such as the design of different questionnaires for different categories of respondents. The design of different questionnaires for different groups of respondents is illustrated in Figure VI-F-1.

Figure VI-F-1

Example of Bradley Fighting Vehicle Questionnaire for Multiple Groups

Questionnaires Designed for:

1. Driver
2. Track commander
3. Gunner
4. Other personnel

2. Order of Different Groups of Questions

a. There is usually a psychological or logical order in which to ask questions, so that the questionnaire flows smoothly from one topic to the next and the respondent is not shifted frequently from one topic to another and back again. However, when a shift is made from one topic to another, it should be apparent to the respondent.

b. It is usually recommended that more difficult or more sensitive questions be asked later in the questionnaire, possibly at the end.

c. One or more easy, non-threatening questions should probably be asked first to build rapport. They should be short and easy to understand and to answer. But they should not be irrelevant to the objectives of the questionnaire. Verbal efforts to build rapport by the questionnaire administrator seem preferable to using questionnaire content to accomplish this task.

G. Number of Response Alternatives

The following sections consider number of response alternatives to use in multiple choice, rating scale, and forced choice items: Section VI-C-3 - formulation of response alternatives; Section VI-H - order of response alternatives; Chapter VII - response anchoring; Chapter VIII - order of perceived favorableness of words and phrases.

One of the basic issues in the use of rating questions or attitude scales is the determination of the optimum number of responses, alternatives or categories. In questionnaire construction, researchers have investigated the utility of having a scale with a greater or smaller number of scale points. Over the years, there have been diverse recommendations on the proper number of scale points or categories to use in questionnaire construction. Investigations have indicated that reliability was optimum for scale points of 2, 5, 10, 11, 20, and 25. Some recent research has proposed the use of a range of scale points between 2 and 10. The reason for concern with the number of response alternatives is due to the belief that a "coarse" scale with too few response alternatives may result in a loss of information concerning the respondents' discrimination powers. It may reduce the respondents' cooperation in rating, as a coarse scale "forces" judgments and thereby irritates some respondents. An extremely "fine" scale, with too many response alternatives, may go beyond the respondents' powers of discrimination, be excessively time consuming, or difficult to score.

1. Number of Response Alternatives with Multiple Choice Items

No firm rules can be established regarding the number of response alternatives to use with multiple choice items. It depends in a large part upon the question being asked, and the number of answers logically possible. The following considerations, however, may be noted:

a. There is some evidence that dichotomous items (items with only two response alternatives) are statistically inferior to items with more than two response alternatives.

b. Dichotomous items are easier to score than nondichotomous items, but they may not be accepted as well by the respondent.

c. A good nondichotomous multiple choice item usually cannot be written as a set of separate dichotomous items.

d. Consideration should be given to the prospect that many response alternatives may make a questionnaire unduly time consuming.

e. The number of choices logically possible or desirable should constitute an upper limit on the number of response alternatives used for an item.

f. Non-existent response alternatives may be checked by the respondent if an answer sheet is used which has more spaces than there are alternative answers; e.g., the answer sheet has five spaces for each question, but some questions have fewer than five alternatives.

2. Number of Response Alternatives with Rating Scale Items

Authorities in psychometrics contend that the optimal number of response alternatives to employ with rating scales is a matter for empirical determination in any situation. They also suggest that considerable variation in number around the optimal number changes reliability very little. These conclusions seem to be supported by the available research literature. Although rules regarding the number of response alternatives to use with rating scales cannot, therefore, be firmly established, the following issues can be considered.

a. The effects of increasing or decreasing the number of response alternatives for a question cannot be generally specified with certainty. Increasing the number of response alternatives does not necessarily increase reliability, and there is no consistent relationship between the number of response alternatives and validity.

b. J. P. Guilford (in Psychometric methods. New York: McGraw-Hill, 1954) reported that seven response alternatives is usually lower than optimal, and it may pay in some favorable situations to use up to 25 scale divisions. Others believe that seven steps or five is optimal. Some believe that five should be used for single or unipolar (one direction)) scales, nine for double or bipolar scales. Many practitioners consistently use five-point scales. Sometimes a nine-point hedonic (pleasure) scale is recommended for food items, and a six-point scale for other uses.

c. The number of response alternatives to use is often determined on the basis of the degree of discrimination required. For example, a nine-point scale may sometimes (but not always) give greater discrimination than a three-point scale. Increases in reliability tend to level off after seven scale points, and there is no apparent advantage in using a large number of scale points.

d. Psychologists with considerable experiences in military operational field testing feel that anything more than five alternatives is too great a number for many junior enlisted personnel to discriminate among. More nonresponses are obtained, and the discrimination power of answered items is not increased.

e. Questionnaire administration time is probably a function of the number of response alternatives.

f. There is some evidence that increasing the number of response alternatives seems to decrease the number of nonresponses and uncertain responses (e.g., "Cannot decide").

g. In addition to the response alternatives representing the rating scale continuum, it may be necessary to add alternatives such as "No opinion" or "Did not experience."

h. Scoring and data analysis considerations may affect the selection of the number of response alternatives. If Chi square tests are sufficient, two or three response alternatives might be adequate. However, if two or three response alternatives are used when nonparametric rank order correlations are employed, substantial "ties" on ranks will result. If parametric statistics are to be employed, more alternatives are usually better, because of the assumption of continuous distributions or interval scale properties.

i. In some situations, fully-labeled scales may discriminate better than only end-anchored scales. Responses to fully-labeled scales may be less skewed than responses to only end-anchored scales.

3. Number of Response Alternatives with Forced Choice Items

A number of different forced choice item formats have been used, such as the following:

a. Two phrases or statements per item, both favorable or both unfavorable, choose the more descriptive or the least descriptive.

b. Three statements per item, all favorable or unfavorable, choose the most and least descriptive statements in each item.

c. Four statements per item, all favorable, choose the two most descriptive statements.

d. Four statements per item, all favorable, choose the most and least descriptive statements.

e. Four statements per item, two favorable and two unfavorable, choose the most and least descriptive statements.

f. Five statements per item, two of which were favorable, one neutral, and two unfavorable in appearance, choose the most and least descriptive.

The evidence is not clear, but three or four statements per item may be preferable to two. One study concluded that the format described in "c" above was superior to the others. It was most bias resistant, yielded consistently high validities under various conditions, had adequate reliability, and was one of the best received by respondents.

H. Order of Response Alternatives

1. General Considerations

The experimental evidence on the effect that the order of presentation of response alternatives for a question has on a subject's choice of response is inconclusive and contradictory. Varying conclusions include:

a. Respondents have a tendency to select the first response alternative in a set more than the others.

b. With multiple choice questions, there is a tendency to choose answers from the middle of the list, if the list consists of numbers. Answers were selected from either the top or bottom of the list, if the alternatives were fairly lengthy expressions of ideas.

c. Longer items produced responses that were closer to the center of the response scale. Shorter items yielded more positive responses.

d. Poorly motivated respondents tend to select the center or neutral alternatives with rating scale items.

e. Fully-labeled response alternatives yielded less skewed response distributions than only labeling the endpoints.

f. On items about which respondents feel strongly, the order of alternatives makes no difference. On items about which the respondent does not feel strongly, most will tend to check the first alternative.

g. Items that were positively worded received higher mean responses than negatively-worded responses.

h. The positive pole of rating scale response alternatives should be presented first since this will improve the reliability of the responses. However, it is important to realize that reliability may increase while validity decreases.

i. Placement of either the positive or negative endpoint at the left-hand side of the semantic differential scale was not associated with response style.

j. Semantic differential scales were found to confound trait self-descriptions with socially desirable responses on clinical instruments. When a socially undesirable adjective anchor was presented first, subjects had a tendency to select adjectives which were opposite in desirability.

Test item form biases are discussed in Section XII-B.

2. Suggested Order for Multiple Choice Items

The following suggestions are offered regarding the order of multiple choice items:

a. When the response alternatives have an immediate apparent logical order (e.g., they all relate to time), they should be put in that order.

b. When the response alternatives are numerical values, they should in general be put in either ascending or decreasing order.

c. When the response alternatives have no immediately apparent logical order, they should generally be put in random order.

d. Alternatives such as "none of the above" or "All of the above" should always be in the last position.

e. Alternate questionnaire forms (e.g., where the order of alternatives is reversed on half of the forms) are often desirable.

f. More abstract types of questions minimize order effects by developing questions which are specific in content instead of general,

3. Suggested Order of Rating Scale Items

Since rating scales call for the assignment of objects along an assumed continuum or in ordered categories along the continuum, it follows that the response alternatives must be in order from "high" to "low" or "low" to "high," with the choice of words for "high" and "low" (the endpoint labels) depending upon the continuum being used. For example, for the continuum satisfactory-unsatisfactory, item 1 in Figure VI-H-1 uses the "high" to "low" order, while item 2 uses the order "low" to "high."

Figure VI-H-1

Example of Rating Scale Item
with Alternate Ordering of Response Alternatives

1. The M16 rifle is:

_____ very satisfactory.

_____ satisfactory.

_____ borderline.

_____ unsatisfactory.

_____ very unsatisfactory.

2. The M16 rifle is:

_____ very unsatisfactory.

_____ unsatisfactory.

_____ borderline.

_____ satisfactory.

_____ very satisfactory.

Many practitioners use the "high" to "low" order. If one has reason to believe that the order of the response alternatives makes a difference, or wishes to make certain that they do not, then the use of alternate questionnaire forms is recommended. Each alternate form should list the response alternatives in a different order. The "good" or "high" end of the scales should be at the same end of each scale for all items in a given questionnaire form, but the order should normally be reversed on 50% of the forms. For example, the order shown in item 1 in Figure VI-H-1 would be used on half of the forms; the order shown in item 2 on the other half. (Normally, there would be only two questionnaire forms, one with each order, but at times alternate forms are also needed for other purposes. Hence, there may be more than two.)

Chapter VII: Response Anchoring

A. Overview

This chapter addresses the "anchoring" of rating scale responses; that is, the words used to define some or all of the response alternatives. Section VII-B shows various types of response anchors, while Section VII-C discusses anchored versus unanchored scales. The amount of verbal anchoring is the topic of Section VII-D, while some procedures for the selection of verbal scale anchors are presented in Section VII-E. Finally, Section VII-F discusses balanced versus unbalanced scales.

It should be noted that Section VI-C 3 discussed the formulation of response alternatives, while the number and order of response alternatives are the topics of Sections VI-G and VI-H, respectively. The order of perceived favorableness of words and phrases is discussed in Chapter VIII.

B. Types of Response Anchors

There are a number of different types of response anchors that can be used with rating scale items. Some have been shown as examples in other chapters, such as Section VI-D. Nine types of response anchors are shown in Figure VII-B-1. The first shows the original form of the semantic differential. It is a combination graphic and verbal scale. Respondents were instructed to place an "X" at a place on the line that would represent their attitude. The use of verbal anchors with a -5 through +5 numerical continuum is shown in item 2 of Figure VII-B-1. Item 3 shows verbal anchors used with a 1 through 11 numerical continuum. There is evidence that variables studied by behavioral scientists are continuously distributed, even though the measuring instruments yield discrete scores. These scores are approximations of the supposedly continuous variables. A combination verbal and numerical continuum (series) is shown in item 4, while a verbal and alphabetical continuum is shown in item 5. Item 6 is similar to item 5 since it too is a verbal continuum. This item lacks the alphabetical and numerical response anchors associated with other verbal anchors. Item 7 is a typical Likert rating scale that calls for a verbal rating to a directional statement that may be phrased either positively or negatively. An example might be "The Modern Volunteer Army places too much emphasis on extrinsic factors (such as beer in the barracks) as opposed to intrinsic, job related factors (such as pay or supervision)." Item 8 is constructed on a continuous scale to obtain more discrimination along the scale line, and it is verbally anchored. Item 9 is one behavioral anchor from a set of nine scale points. This particular behavioral anchor has a scale point of four.

Conflicting empirical evidence exists regarding the reliability of scales with verbal anchors and verbal response alternatives so that neither is superior to that of purely numerical scales. Some feel that adding verbal anchors to a scale will increase reliability. Recent research in ergonomics and other related applications indicates that either numerical response alternatives or verbal response alternatives are psychometrically acceptable. If verbal anchors are used, be sure they are properly developed.

Figure VII-B-1

Types of Response Anchors

1. Combination graphic and verbal scale.

Strong _____: _____: _____: _____: _____: _____: _____: _____: _____ Weak

Extremely Quite Slight Slight Quite Extremely

LOW HIGH

2. Verbal anchors with a -5 through +5 numerical continuum (series).

Definitely dislike ... Definitely like

-5 -4 -3 -2 -1 0 +1 +2 +3 +4 +5

3. Verbal anchors with a 1 through 11 numerical continuum (series).

Definitely dislike ... Definitely like

1 2 3 4 5 6 7 8 9 10 11

4. A verbal and numerical continuum (series).

Dislike completely	Dislike somewhat	Dislike a little	Neither like nor dislike	Like a little	Like somewhat	Like completely
1	2	3	4	5	6	7

5. A verbal and alphabetical continuum (series).

Below Average	Average	Above Average	Well Above Average	Outstanding
(A)	(B)	(C)	(D)	(E)

Figure VII-B-1 (Cont.)

Types of Response Anchors

6. A verbal continuum (series).

Below average	About average	A little better	A lot better	One of the best	None better

7. A verbal continuum (series). (Likert rating scale)

__Agree strongly __Agree __Undecided __Disagree __Disagree strongly

8. Combination verbal and continuous (series) scale.

Attribute

negative neutral positive

9. Combination behavioral anchor and numerical scale point.

Scale Point	
4	Many troops in this unit would leave the post as quickly as possible after duty hours to avoid doing any extra work.

C. Anchored Versus Unanchored Scales

A number of studies have been conducted on the topic known as "anchoring effects." Unfortunately, the research evidence is contradictory as to whether anchored or unanchored scales should be used. It has been noted that unanchored scales may well be anchored by the question stem, so that the response alternatives may not have to be. When only one end of a scale is anchored, some studies have found a tendency for respondents to move toward that extreme. But other studies have found the opposite tendency. At least one study found that judgment and response time is decreased with anchoring. In practice, then, it is usually best to use anchored scales.

D. Amount of Verbal Anchoring

Obviously, the amount of verbal anchoring of a rating scale item can vary. It can be anchored at the center, or on the ends, or both, or at many points on the entire continuum. There is some evidence that more descriptive data can be obtained with more anchoring, and that greater scale reliability is achieved with added verbal anchoring. In one study, scales labeled at only the extreme endpoints resulted in responses that were skewed toward the positive end of the scale. Scales with verbal descriptors for all response alternatives may also be better predictors of behavior. On the other hand, adding examples to definitions does not seem to help much. (See also Section VI-G regarding the number of response alternatives to employ.) Fully labeled scale points may encourage a more balanced response distribution.

E. Procedures for the Selection of Verbal Scale Anchors

Some guidance can be offered regarding the selection of verbal scale anchors. See also Chapter VIII.

1. Pretests for the selection of verbal anchors are valuable in building scale content. Rather than employing anchors which seem appropriate, anchors should preferably be selected by respondents similar to those who will be participating in the study.

2. Scale endpoints that are unrealistically extreme, such that few if any respondents would select them, should be avoided. For example, it may be seldom that "Never" or "Always" apply. The use of "Rarely" and "Usually" may be more appropriate. There are instances, however, where extreme statements are realistic. The decision here often requires experience with what is being rated.

3. Analysis of data is normally facilitated if verbal scale anchors selected for rating scales are of equal distance from each other in terms of scale values. See, however, Chapter VIII.

4. Scales can be anchored by examples of expected behavior based upon observations of behavior. There are a wide variety of behavioral scales using variations of the Smith and Kendall format. These scales use behavioral anchors constructed from critical incidents. Procedures for establishing behavioral anchors may be found in the following references.

   a. Bernardin, H. J., La Shells, M. B., Smith, P. C., & Alvares, K. M. (1976, February). Behavioral expectation scales: Effects of developmental procedures and formats. Journal of Applied Psychology, 61(1), 75-79.

   b. Borman, W. C., & Dunnette, M. (1975). Behavior-based versus task-oriented performance ratings: An empirical study. Journal of Applied Psychology, 60, 561-565.

   c. Finley, D. M., Osborn, H. G., Dubin, J. A., & Jeanneret, P. R. (1977). Behaviorally based rating scales: Effects of specific anchors and disguised scale continua. Personnel Psychology, 30, 659-669.

   d. Fivars, G. (1975). The critical incident technique: A bibliography. JSAS Catalog of Selected Documents in Psychology, 5, 210.

   e. Landy, F. J., & Barnes, J. L. (1979). Scaling behavioral anchors. Applied Psychological Measurement, 3(2), 193-200.

   f. Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

F. Scale Balance, Midpoints, and Polarity

1. Balanced Versus Unbalanced Scales

Historically, balanced scales have been preferred by researchers. A scale is balanced when it has a number of positive response alternatives equal to the number of negative alternatives, regardless of the presence or absence of an "indifferent," neutral, or mid-scale category. A "Don't know" response alternative, if present, is not considered to be part of the scale, so is not counted when deciding if the scale is balanced. See the examples of balanced and unbalanced scales in Figure VII-F-1. Unbalanced scales may be employed if pretest results indicate that many respondents will be choosing extreme response alternatives at one end of a scale, producing a skewed distribution of responses rather than the statistically expected normal distribution around the mean attitude. To reduce the piling up of responses at one end of a scale, - or, to add to your ability to discriminate among responses in that region - the scale is made unbalanced by adding more response alternatives on the side of the scale where the piling is likely to occur. This practice tends to spread the distribution of responses more evenly along the scale continuum.

In cases where one has no advance information or other basis for expecting responses to be largely one-sided, it is normally desirable to have an equal number of positive and negative response alternatives; i.e., a balanced scale.

2. Midpoints

Scales may or may not include a midpoint or mid-scale response alternative. This does not affect their classification, but does affect their response distributions. There is research evidence that when a middle position is offered on a scale, there is a shift in the distribution of ratings. Up to 10-20% or more of the ratings may shift into the midpoint causing a decline in the polar positions. Even so, questionnaires that have response distributions that include a midpoint yield similar distributions to those without the midpoint. The inclusion or exclusion of a midpoint probably won't influence the response distribution that much as long as there are at least five scale points.

As examples, items 1c, 2a, and 3 in Figure VII-F-1 show scales with no mid-scale point. One might exclude the mid-scale point for items where it is judged that respondents ought to have a sufficient basis for being pro or con, and where one desires to force respondents away from an "on the fence" position. Bipolar scales should be balanced in terms of the degree of extremeness denoted by the endpoint anchors. For example, if "Never" is used, then "Always" should be used as the opposite endpoint.

Figure VII-F-1

Examples of Scale Balance, Midpoints, and Polarity

1. Balanced bipolar scales.

a.
Very progressive
Progressive
Moderately progressive
Neither progressive nor conservative
Conservative
Very conservative

b.
Effective
Fairly effective
Borderline
Fairly ineffective
Ineffective

c.
Very effective
Somewhat effective
Somewhat ineffective
Very ineffective

d.
Very satisfied
Satisfied
Borderline
Dissatisfied
Very dissatisfied

2. Unbalanced bipolar scales.

a.
Enthusiastic
Extremely favorable
Very favorable
Favorable
Fair
Poor

b.
Quite good
Rather good
Somewhat poor
Rather poor
Quite poor
Very poor

3. Unbalanced Scale (unipolar).

Very much
Much
Some
A little
None

3. Polarity

Scales may be bipolar or unipolar. Item 3 in Figure VII-F-1 illustrates a unipolar scale. Its basic feature is that it represents the thing being assessed as having from none to a maximum - with n steps in between - of some property. The question of balance only arises for bipolar scales. Many a bipolar scale could be redesigned as a unipolar scale. Instead of item 1c in Figure VII-F-1, one's question about effectiveness (not given) could have been followed by this unipolar scale of effectiveness: maximum effectiveness, great effectiveness, moderate effectiveness, slight effectiveness, and no effectiveness.

Semantic preferences may determine whether the questionnaire writer uses bipolar or unipolar scales.

Chapter VIII: Empirical Bases for Selecting Modifiers for Response Alternatives

A. Overview

When constructing a questionnaire, it is often necessary to select adjectives, adverbs, or adjective phrases to use as response alternatives. The words selected for response alternatives should be clearly understood by the respondents to the questionnaire. and they should have precise meaning. There should be no confusion among respondents as to whether one term denotes a higher degree of favorableness or unfavorableness than another.

There is no need to guess which phrases or words are the best to use as response alternatives. Many studies have been conducted in order to determine the perceived favorableness of commonly used words and phrases. These studies have determined scale values and variances for words and phrases which can be used to order the response alternatives. In some of the studies, ambiguous words and words that are not appropriate to use as response alternatives have been identified.

The results of these studies and the experience of questionnaire designers have been incorporated into this chapter in order to offer guidelines and suggestions to be used in selecting response alternatives. This chapter includes lists of words and procedures to use in selecting response alternatives. Many lists of phrases with mean scale values and standard deviations are presented. The scale values are given for the purpose of selecting response alternatives, not for the purpose of assigning scale values to response alternatives for data analysis purposes.

Section VIII-B discusses things to consider in selecting response alternatives; Section VIII-C covers the selection of response alternatives denoting degrees of frequency; Section VIII-D, the selection of response alternatives using order of merit lists of descriptor terms; Section VIII-E, the selection of response alternatives using scale values and standard deviations. Section VIII-F includes sample sets of response alternatives.

Scale values, standard deviations, and interquartile ranges reported in this chapter have been taken from data presented in the following studies:

1. Altemeyer, R. A. (1970). Adverbs and intervals: A study of Likert scales. Proceedings of the Annual Convention of the American Psychological Association, 5(pt. 1), 397-398.

2. Backstrom, C. H., & Hurchur-Cesar, G. (1981). Survey research. New York, NY: John Wiley & Sons.

3. Beltramini, R. F. (1982). Rating-scale variations and discriminability. Psychological Reports, 50, 299-302.

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

5. Boote, A. S. (1981). Reliability testing of psychographic scales. Journal of Advertising Research, 21(5), 53-60.

6. Cliff, N. (1959). Adverbs as multipliers. Psychological Review, 66, 27-44.

7. Dodd, S. C., & Gerberick, T. R. (1960). Word scales for degrees of opinion. Language and Speech, 3, 18-31.

8. Dolch, N. A. (1980). Attitude measurement by semantic differential on a bipolar scale. The Journal of Psychology, 105, 151-154.

9. Gividen, G. M. (1973, February). Order of merit- descriptive phrases for questionnaires. Unpublished report, available from the ARI Field Unit at Fort Hood, TX.

10. Innes, J. M. (1977). Extremity and "don't know" sets in questionnaire response. British Journal of Social Clinical Psychology, 16, 9-12.

11. Ivancevich, J. M. (1980). Behavioral expectation scales versus nonanchored and trait rating systems: A sales personnel application. Applied Psychological Measurement, 4(1), 131-133.

12. Jones, L. V., & Thurstone, L. L. (1955). The psychophysics of semantics: An experimental investigation. Journal of Applied Psychology, 39, 31-36.

13. Mathews, J. L., Wright, C. E., & Yudowitch, K. (1975, March). Analysis of the results of the administration of three sets of descriptive phrases. Palo Alto, CA: Operations Research Associates.

14. Mathews, J. L., Wright, C. E., Yudowitch, K. L., Geddie, J. C., & Palmer, R. L. (1978, August). The perceived favorableness of selected scale anchors and response alternatives (Technical Paper 319). Palo Alto, CA: Operations Research Associates, and Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (DTIC No. AD A061755)

15. Menezes, D., & Elbert, N. F. (1979). Alternative semantic scaling formats for measuring store image: An evaluation. Journal of Marketing Research, 16(1), 80-87.

16. Mosier, C. I. (1941). A psychometric study of meaning. Journal of Social Psychology, 13, 123-140.

17. Myers, J. H., & Warner, W. G. (1968). Semantic properties of selected evaluation adjectives. Journal of Marketing Research, 5, 409-412.

18. Presser, S., & Schuman, H. (1980). The measurement of a middle position in attitude surveys. Public Opinion Quarterly, 44(1), 70-85.

19. Reynolds, T. J., & Jolly, J. P. (1980). Measuring personal values: An evaluation of alternative methods. Journal of Marketing Research, 17, 531-536.

20. Schuman, H., & Presser, S. (1981). Questions and answers in attitude surveys: Experiments on question form, wording, and context. New York: Academic Press, Inc.

21. U.S. Army Test and Evaluation Command (1973). Development of a guide and checklist for human factors evaluation of Army equipment and systems. U.S. Army Test and Evaluation Command (TECOM).

B. General Considerations in the Selection of Response Alternatives

There are several ways of selecting response alternatives. These ways are dependent on the purpose of the questionnaires and/or on the way the data will be analyzed. There are specific considerations when selecting response alternatives for balanced scales, when selecting response alternatives with extreme values, and when developing equal interval scales. There are also general things to consider in the selection of any response alternative.

In some cases, it is desirable to select response alternatives on more than one basis. For example, mutually exclusive phrases may be selected also on the basis of parallel wording.

1. Matching the Question Stem

Descriptors should be selected to follow the question stem. For example, if the stem asks for degrees of usefulness, descriptors such as "Very useful" and "Of significant use" should be used. In some cases, this may mean rewording the question stem so that appropriate response alternatives can be selected.

2. Mixing Descriptors

Descriptors on different continuums should usually not be mixed. For example, "Average" should never be used with quantitative terms or qualitative terms such as "Excellent" or "Good" (since "average" performance for a group may very well be excellent or good or even poor). If the descriptors are selected for use with a question stem asking about satisfactory or unsatisfactory, the word "Satisfactory" or "Unsatisfactory" (or a synonym) should normally be in every response alternative, except perhaps for a neutral response alternative.

Some experts go as far as to say that the wording of the response alternatives should be parallel for balanced scales. For example, if the phrase "Strongly agree" is used, then the phrase "Strongly disagree" should also be used. By reviewing some of the studies that have determined scale values for descriptors, it can be seen that some pairs of parallel phrases are not equally distant from a neutral point or from other phrases in terms of their scale values. Hence, parallel wording may not always provide equally distant pro and con response alternatives, although they may be perceived as symmetrical opposites.

Using descriptors from one continuum or descriptors with parallel wording for a given questionnaire item has advantages. The advantages are that the response alternatives will usually fit the stem better, and they will be parallel to each other in meaning and appearance.

3. Selecting Response Alternatives with Clear Meaning

Some words are difficult for respondents to use in answering questions. This difficulty may be the result of the respondent being ignorant of the meaning of the word, or not being able to rate the word in terms of degrees on specific scales. Such words should not be used as response alternatives. Some studies asked the respondents to indicate which words they were unable to rate. Table VIII-B-1 lists examples of words that were unrateable by subjects.

Table VIII-B-1

Words Considered Unrateable by Subjects

Phrase	Phrase
Adverse	Noxious
Appalling	Peerless
Base	Satiating
Despicable	Seemly
Expedient	Superlative
Fit	

From: Mosier 1941a.

Some words appear to have two or more distinct meanings. When these words are rated on a continuum of favorableness-unfavorableness, many respondents will mark one part of the scale, while the other respondents will mark a different place on the scale. This depends on how they interpret the words, and may result in a bimodal distribution. Such words also should not be used as response alternatives. A list of words evoking bimodality of response is given in Table VIII-B-2.

Table VIII-B-2

Words Evoking Bimodality of Response

Word(s)	Word(s)
Acceptable	Irresistable
Amazing	Normal
Bearable	Tempting
Completely indifferent	Unfit
Extremely indifferent	Unspeakable
Highly indifferent	Unusually indifferent
Important	Very indifferent
Indifferent	Very, very indifferent
Indispensable	

From: Mosier 1941a

4. Selecting Nonambiguous Terms/Descriptors

Some descriptors are more ambiguous than others. The more ambiguous the descriptor, the more varied the respondents' interpretations of the degree of favorableness denoted by the descriptor. The ambiguousness of a descriptor is measured by the variability of responses given to the item. One measure of variability is the standard deviation. When available, standard deviations (SD) are given with scale values in this chapter. Another measure used to show variability is the interquartile range. This measure is indicated in this chapter with scale values only when the standard deviations were unavailable.

It is most desirable to select terms with small ranges or small standard deviations, as they will have less ambiguous meaning to respondents. Also, selecting a term with a small standard deviation decreases the chances of the meaning of the term overlapping with the meaning of neighboring terms.

5. Selecting Response Alternatives

When balanced scales with two, three, four, or five descriptors are sufficient for describing the distribution of respondents' attitudes or evaluations, the questionnaire writer can compose them quite satisfactorily by using a term and its literal opposite (effective vs. ineffective; pleasing vs. unpleasing) for two of the terms. A more extreme pair can be produced by using "Very" to modify these two terms.

The first of several intended studies of how people rate/order terms that might be used for rating scale descriptors was conducted by Operations Research Associates and ARI just prior to the writing of this manual. Its results may assist questionnaire developers who need unbalanced scales or scales with more than five descriptors. In the study, each of 100 Army personnel was asked to assign a scale value ranging from -5 (most negative) to +5 (most positive) to each term in three different sets of terms, totaling over 100 descriptors.

Tables VIII-B-3 and VIII-B-4 give samples of descriptors from this study for which mean scale values and standard deviations have been calculated. The list in Table VIII-B-3 was derived by first selecting the descriptor with the largest positive mean. The next descriptor selected has a mean that is at least one standard deviation lower. The implication of the gap of one standard deviation is that not more than 16% of the people would have assigned a lower scale value to the first descriptor than they did to the second descriptor, and vice versa. To this extent, the raters disagreed on the ordering of these two terms when rating about 50. The third descriptor on the list has a mean scale value yet another standard deviation lower. This process was repeated until the descriptor with the lowest mean scale value was selected. A descriptor was not used if its standard deviation was greater than 1.000.

The list on Table VIII-B-4 was constructed again by skipping at least one standard deviation between adjacent terms; however, the starting point was at the middle, with the word "Neutral."

Use of Table VIII-B-3 as a 10-descriptor unbalanced scale is not highly recommended. If one wanted a nine-descriptor scale, one could use the four adverbs appearing in front of "Acceptable" in the table in that same location, and also use them in front of "Unacceptable" in reverse order to create a semantically balanced and ordered scale. Or, one could use the five adverbs, now shown below "Neutral," both above and below "Neutral" to create an 11-descriptor scale of acceptability (or effectiveness, or satisfactoriness, etc.). "Neutral," however, may not be a suitable midpoint term here as the respondent who has neutral feelings (i.e., does not know or does not care) might check this response, whereas the term "Neutral" is intended to specify, for example, a midpoint between "barely acceptable" and "barely unacceptable."

Table VIII-B-3

Sample List of Phrases
Denoting Degrees of Acceptability

Phrase	Mean	SD
Wholly acceptable	4.725	.563
Highly acceptable	4.040	.631
Reasonably acceptable	2.294	.722
Barely acceptable	1.078	.518
Neutral	.000	.000
Barely unacceptable	-1.100	.300
Rather unacceptable	-2.020	.836
Substantially unacceptable	-3.235	.899
Highly unacceptable	-4.220	.576
Completely unacceptable	-4.900	.361

From: Mathews, Wright, and Yudowitch (1975).
See Section VIII-A 13.

Table VIII-B-4

A Second Sample List of Phrases
Denoting Degrees of Acceptability

Phrase	Mean	SD
Very, very acceptable	4.157	.825
Largely acceptable	3.137	.991
Mildly acceptable	1.686	.700
Sort of acceptable	.940	.645
Neutral	.000	.000
Barely unacceptable	-1.100	.300
Rather unacceptable	-2.020	.836
Substantially unacceptable	-3.235	.899
Highly unacceptable	-4.294	.535
Completely unacceptable	-4.900	.361

From: Mathews, Wright, and Yudowitch (1975).
See Section VIII-A 13.

While the scale values from the studies cited are useful, further refinement is possible. That is, once having selected a candidate scale (set of descriptors), one could then conduct another study to determine if relevant judges would assign scale values indicating equal intervals (among means) for the terms on the candidate scale.

6. Selecting Descriptors for Endpoints

Once the decision has been made as to how extreme the endpoints of a scale should be (see Section VII-E 4), the descriptors should be selected accordingly. If extreme endpoints are desired, descriptors that have extreme meanings should be selected. One guideline that can be used in selecting these descriptors is to use those that have the highest and lowest scale values. Another guideline is to review the descriptors in terms of their apparent meanings. If less extreme endpoints are desired, descriptors that do not have extreme scale values and that do not have the apparent extreme meanings should be selected.

There has been conflict about whether fully-labeled scales are psychometrically superior to scales labeled only at the endpoints. Some evidence supports fully-labeled scale points which appear to produce response distributions that are less skewed.

7. Selecting Midpoint Responses

Whether a middle response alternative is included or excluded on a scale won't make that much difference as long as the number of scale points is at least five and not more than eleven. For respondents who have a weak opinion, eliminating the middle alternative will force a response toward either end of a bipolar scale. Including the middle alternative may increase differentiation of response, and may be useful for individuals who have a strong opinion on the topic. Overall, response distributions for scales that include the middle alternative look about the same as the distributions without the middle response alternative. The decline in responses to endpoints of the scale accounts for the shift in response when a middle alternative is offered.

Identification and selection of the label for the middle alternative is dependent on the overall selection of the other response alternatives. Different populations will perceive response alternatives with divergent perceptions. The means and variances for agreement on the semantic meaning of response alternatives will vary by population. To identify appropriate response alternatives, a sample from the target population could rate response alternatives for agreement for semantic meaning. Rating response alternatives by themselves may produce different results than rating response alternatives in conjunction with the item stem.

In selecting a descriptor for a midpoint response, it is necessary to use a descriptor that is neutral (neither positive nor negative) in meaning. Some of the commonly used midpoints do not appear as neutral as might be expected to some respondents.

Table VIII-B-5 lists several candidate midpoint terms with their scale values and standard deviations. This list may be helpful in selecting midpoint responses.

Table VIII-B-5

Candidate Midpoint Terms' Scale Values and Standard Deviations as Determined by Several Different Studies

Term	Mean Scale Value	SD	Theoretical Middle Scale Value
About average	3.77	.85	3.50
Acceptable	.73	.66	.00
Acceptable	11.12	2.59	10.00
Acceptable	2.39	1.46	.00
All right	10.76	1.42	10.00
Average	3.08	--	3.00
Average	.86	1.08	.00
Average	10.84	1.55	10.00
Borderline	-.02	.32	.00
Borderline	.00	.20	.00
Borderline	-.06	.31	.00
Doesn't make any difference	2.83	3.73[a]	5.00
Don't know	4.82	.82[a]	5.00
Fair	6.50	--	5.50
Fair	.78	.85	.00
Fair	9.52	2.06	10.00
Fair	4.96	.77[a]	5.00
Neutral	.00	.00	.00
Neutral	.02	.18	.00
Neutral	9.80	1.50	10.00
Neutral	10.18	2.01	10.00
Normal	6.70	1.43	6.00
Ordinary	6.50	1.43	6.00
O.K.	.87	1.24	.00
O.K.	10.28	1.67	10.00
So-so	10.08	1.87	10.00
Undecided	4.76	3.73[a]	5.00

[a]Interquartile range shown rather than the standard deviation

Words commonly used for midpoint responses are discussed below:

a. Average.

"Average" should never be used in conjunction with adjectives such as "Excellent," "Good," etc. "Average" has no meaning when used with these words. For example, "Average" performance may be superior or it may be completely unsatisfactory. Furthermore, most evaluators do not have the experience or competence to even know what an "Average" performance is. Typically, when "Average" is used on a field test evaluation form, only 5% or 10% of responders rate the subject as below average and 30% or 40% rate it above average. The data from such a question indicate that the response alternatives are not well formulated. Therefore, as a general rule, it is usually inappropriatee to use any term of "Average" in a questionnaire, and it is always inappropriate to use "Average" in conjunction with phrases such as "Excellent," "Good," "Poor," etc.

If "Average" is used, it should be with extreme care and only when one is interested in comparing performances or items with each other. It should not be used when one desires to find out how "good" or how "bad" an item or performance is. Significantly above average performance may be extremely unsatisfactory.

b. No opinion.

"No opinion" is unacceptable as a mid-scale term, as it usually denotes that a person has no opinion due to lack of knowledge or due to not having thought about an issue. "No opinion" can be used as a response alternative if it represents a specific type of information that is wanted.

c. Neutral.

"Neutral" is considered as a less desirable mid-scale term to use than "Borderline." Although every respondent in the study gave the term zero, the meaning on a questionnaire is not clear (see page VIII-B 4). Two out of 52 respondents indicated it was unrateable. In another study, "Neutral" had a mean scale value of .02 and a standard deviation of .18. Because of the ambiguity of meaning of "Neutral" (e.g., feeling of the respondent versus midpoint alternative), it is not recommended that it be used as midpoint on most questionnaires.

d. Marginal.

"Marginal" is sometimes used as a midpoint response alternative. Interviews with test subjects indicated that the term "Marginal" in most cases had a meaning of above "Borderline" or still satisfactory, but very close to being unsatisfactory. Hence, indications are that there may be more desirable terms to use than "Marginal."

e. Borderline.

"Borderline" is preferred by some experts as a midpoint response. In an administration to Fort Hood soldiers of over 1,500 questionnaires using the term "Borderline" as a midpoint, there was not one instance of reported confusion among those completing the questionnaires. However, there were times when "Borderline" had a larger standard deviation than "Neutral." (Again, "Neutral" by definition implies zero to most persons, but its frame of reference is ambiguous).

f. Uncertain.

"Uncertain" is unacceptable as a midpoint term, as it implies that with additional knowledge or thought a decision could be made that would fall into one of the other categories.

g. Undecided.

"Undecided" is also unacceptable as a mid-scale term for the same reasons as "Uncertain."

h. Neither agree nor disagree.

"Neither agree nor disagree" and similar descriptors written in this form may be used as midpoint responses. They have the advantage of paralleling the rest of the descriptors in the set, and they denote a position exactly in the middle of the endpoints. This term, like "Neutral," can also imply uncertainty, indecision, or a lack of knowledge rather than a firm knowledge that it represents a midpoint.

i. No effect.

"No effect" may be employed as a midpoint term when it is used with a set of descriptors to measure the type of effect that an activity will have. For instance, it can be used on a continuum from beneficial to detrimental.

j. Ordinary.

"Ordinary" should not be used as a mid-scale term. In one study, use of the term "Ordinary" as the mid-scale value resulted in a marked skewing of responses at the low end of the scale. This resulted from the common use of "Ordinary" to imply inferiority.

k. Fair.

"Fair" should not be used as a mid-scale term. In one study, the median scale value for "Fair" was a full point above the mid-scale point. It appears for some subjects that the meaning of "Fair" is distinctly favorable.

l. Acceptable.

"Acceptable" is not a desirable word to use as a mid-scale item. In one study it exhibited a marked bimodality of response, indicating that subjects disagreed on the degree of favorableness noted by the term. In a recent study, "Acceptable" had a large standard deviation of 1.46.

m Normal.

"Normal" is not a desirable word to use as a mid-scale item. In one study it exhibited a marked bimodality of response, indicating that the word "Normal" has different meanings for different subjects. This term would be classified as a synonym for "Average."

n. Medium.

"Medium" may possibly be used as a midpoint term. In one study there was a piling up of judgments for "Medium" at the middle scale position.

o. O.K. or all right.

"O.K." or "All right" have been used sometimes as midpoint response alternatives. However, they have a tendency to be rated somewhat positively. They also have larger standard deviations than other terms mentioned, indicating that there is ambiguity in their meaning.

p. So-so.

"So-so" is another term sometimes used as a midpoint response. In one study it had a scale value of 10.08, which was very close to the middle scale value of 10.00; but it also had a fairly large standard deviation of 1.87. Its use is not recommended.

q. Don't know.

"Don't know" is an unacceptable term to use as a midpoint. It usually means to the subject that, with additional knowledge or more time to think about the issue, he/she could choose one of the other alternatives.

r. Doesn't make any difference.

"Doesn't make any difference" should not be used as a midpoint response alternative because it implies a more negative value than a middle value. In one study it had a scale value of 2.83, where the middle scale value was 5.00. It also had an interquartile range of 3.13, which means that there was a lot of disagreement among subjects as to its meaning.

8. Selecting the Don't Know Response Alternative

Sometimes respondents are known for their tendency to mark the "Don't know" category. This selection is made when they are not aware of the content in a question or when they refuse to express their opinion. Researchers are not able to predict who would make a shift into a "Don't know" category when it is offered. It has been a common practice in public opinion survey research to leave out the "Don't know" category. Individuals who volunteer a "Don't know" response (even though it is not offered) will have it included as their selection of a response alternative. Human factors researchers who target surveys toward respondents who may not have access to specific experiences or equipment would appropriately include the "Don't know" category. When included in a survey, the "Don't know" category should be set apart from the other response alternatives to avoid confusing it with other categories. An example of the "Don't know" response alternatives is presented in Figure VIII-B-1.

Figure VIII-B-1

Inclusion of the "Don't Know" Response Alternative for a Maintenance Vehicle Questionnaire

Ease of Use Rating Scale

5	4	3	2	1	DK
Very Easy	Easy	Borderline	Difficult	Very Difficult	Don't Know

How easily can you:

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

9. Selecting Positive and Negative Descriptors

If a balanced scale is desired, it is necessary to select an equal number of positive and negative descriptors. In most cases, it is easy to determine if a descriptor is positive or negative by seeing on which side of the (zero) midpoint its scale value falls. For example, "Mildly like" has a positive scale value, and "Mildly dislike" has a negative scale value.

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

The presentation of a positive or negative endpoint displayed first at the left-hand side of the scale has been investigated. It was found that order of presentation for the placement of positive or negative endpoints was not associated with response style (non-trait measures). Measures of personality traits are most influenced by balancing positive and negative descriptors or stems. Operational test and evaluation survey constructors would not need to be concerned about positioning the positive or negative endpoint first.

10. Selecting Type of Response Alternative

Points along the continuum of a scale have been anchored by many different types of response alternatives. For example, there have been response alternatives such as numbers, adjectives, adverbs, phrases, sentences, descriptions of behavior, etc. It does not seem to matter if response alternatives are numerical, verbal (one word), phrases, or behavioral. No one type of response alternative has proven superior to another.

11. Selecting Terms Showing Equal Intervals

Some experts argue that, in order to perform analyses on the basis of numerical values or weights, the intervals between rating scale response alternatives should be equal. This would be desirable, but in many cases it is impossible because many words have not been assigned scale values. But when scale values are available, the response alternatives can be selected as equally distant apart as possible when doing so is considered important.

There is a tendency for some questionnaire constructors to select phrases with parallel wording to indicate equal intervals. (They may also do so for other reasons.) However, if equal intervals are considered important, phrases should be selected based upon scale values if available. For example, in Table VIII-E-9, "Highly adequate" has a scale value of 3.843 while the parallel term "Highly inadequate" has a scale value of -4.196. This places "Highly inadequate" further away from the neutral point than "Highly adequate."

12. Use of Unscaled Terms

Some discussion is in order regarding the use of terms ignoring their scale values or to which no scale values have been assigned. An illustration of the first of these practices is from a study in which ARI had 21 Army officers involved in operational field testing rank-order 16 terms that included "Outstanding," "Superior," "Excellent" and "Very good." "Excellent" was ranked as less positive than "Outstanding" by 14 of the officers, while it was ranked as less positive than "Superior" by 17 of the officers. However, there was maximum disagreement as to whether "Outstanding" or "Superior" was first or second on the scale. That is, 12 rated "Superior" first and "Outstanding" second, while nine of the officers assigned the reverse ordering to these two words. All officers ranked "Outstanding," "Superior," and "Excellent" as more positive than "Very good." "Outstanding" is sometimes interpreted to denote only that the performance is among the best of a group -- without any implication as to quality, e.g., although a student's grade of 65 out of 100 points was failing, his/her performance may have been "Outstanding" since no other student in the class scored above 60!

What are the consequences to the developer of rating scale items of discovering a mean 50%-50% split as in the ordering of "Outstanding" and "Superior?" Does it mean they cannot be used together as part of the descriptors of a rating scale item? The answer is, "Normally yes." In Figure VIII-B-2, we would have better discrimination if "Outstanding" were replaced by "Excellent," with the position formerly occupied by "Excellent" being filled by "Very good." "Superior" and "Outstanding" or similarly overlapping terms should normally not be used on the same scale.

Figure VIII-B-2

Two Formats Using "Outstanding" and "Superior"

1. _____ 1. Superior

_____ 2. Outstanding

_____ 3. Excellent

_____ 4. Good

_____ 5. Fair

_____ 6. Poor

2. Superior Outstanding Excellent Good Fair Poor

(Circle one word)

When functioning as questionnaire consultants or developers in field test situations where respondents are enlisted personnel, ARI has recommended and used very little variety in its rating scale items. Arrays such as those shown in Figure VIII-B-3 are almost always proposed and used. Sometimes the middle term is deleted. Several reasons for the lack of variety are that a standard format facilitates: (1) comparability of rating distributions with previous tests, and (2) understanding by soldier respondents, who are often not high school graduates.

Figure VIII-B-3

Response Alternatives
Frequently Recommended by ARI

( ) Very satisfactory

( ) Satisfactory

( ) Borderline

( ) Unsatisfactory

( ) Very unsatisfactory

( ) Very effective

( ) Effective

( ) Borderline

( ) Ineffective

( ) Very ineffective

( ) Very acceptable

( ) Acceptable

( ) Borderline

( ) Unacceptable

( ) Very unacceptable

C. Selection of Response Alternatives Denoting Degrees of Frequency

Some questionnaire designers use verbal descriptors to denote degrees of frequency. Table VIII-C-1 shows such a list of verbal descriptors. A study showed that there was a great deal of variability in meaning for frequency phrases. Questionnaires should, whenever possible, use response alternatives that include a number designation or percentage of time meant by each word used as a response alternative.

Table VIII-C-1

Degrees of Frequency

Phrase	Scale Value	Inter-Quartile Range
Always	8.99	.52
Without fail	8.89	.61
Often	7.23	1.02
Usually	7.17	1.36
Frequently	6.92	.77
Now and then	4.79	1.40
Sometimes	4.78	1.83
Occasionally	4.13	2.06
Seldom	2.45	1.05
Rarely	2.08	.61
Never	1.00	.50

From: Dodd and Gerberick (1960).
See Section VIII-A 7.

D. Selection of Response Alternatives Using Order of Merit Lists of Descriptor Terms

An order of merit list of descriptors does not provide scale values nor show the variance of each phrase along some continuum. In addition, the list does not represent an equal interval scale. However, such lists are still useful for selecting response alternatives if the main concern is to select response categories so that each respondent will agree on the relative degree of "goodness" of the terms. Tables VIII-D-1 and VIII-D-2 give examples of order of merit lists of descriptor terms.

Table VIII-D-1

Order of Merit of Selected Descriptive Terms

Order of Merit	Descriptive Term
1	Very superior
2	Very outstanding
3	Superior
4	Outstanding
5	Excellent
6	Very good
7	Good
8	Very satisfactory
9	Satisfactory
10	Marginal
11	Borderline
12	Poor
13	Unsatisfactory
14	Bad
15	Very poor
16	Very unsatisfactory
17	Very bad
18	Extremely poor
19	Extremely unsatisfactory
20	Extremely bad

From: Givíden (1973). Section VIII-A 9.

Table VIII-D-2

Order of Merit of Descriptive Terms
Using "Use" as a Descriptor

Order of Merit	Descriptive Term
1	Extremely useful
2	Very useful
3	Of significant use
4	Of considerable use
5	Of much use
6	Of moderate use
7	Of use
8	Of some use
9	Of little use
10	Not very useful
11	Of slight use
12	Of very little use
13	Of no use

From: Gividen (1973). See Section VIII-A 9.

E. Selection of Response Alternatives Using Scale Values and Standard Deviations

Using scale values and standard deviations to select response alternatives will give a more refined set of phrases than using an order of merit list. Other sections above have discussed specific considerations in selecting descriptors. In general, response alternatives selected from lists of phrases with scale values should usually have the following characteristics:

1. The scale values of the terms should be as far apart as possible.

2. The scale values of the terms should be as equally distant as possible.

3. The terms should have small variability (small standard deviations or interquartile ranges).

4. Other things being equal, the terms should have parallel wording.

Tables VIII-E-1 through VIII-E-24 give lists of phrases which have scale values and, when possible, standard deviations or interquartile range. They are based on empirical evidence, and may be used to select response alternatives.

Table VIII-E-1

Acceptability Phrases

Phrase	Average	SD
Excellent	6.27	.54
Perfect in every respect	6.22	.86
Extremely good	5.74	.81
Very good	5.19	.75
Unusually good	5.03	.98
Very good in most respects	4.62	.72
Good	4.25	.90
Moderately good	3.58	.77
Could use some minor changes	3.28	1.09
Not good enough for extreme conditions	3.10	1.30
Not good for rough use	2.72	1.15
Not very good	2.10	.85
Needs major changes	1.97	1.12
Barely acceptable	1.79	.90
Not good enough for general use	1.76	1.21
Better than nothing	1.22	1.08
Poor	1.06	1.11
Very poor	.76	.95
Extremely poor	.36	.76

From: U.S. Army (1973). See Section VIII-A 21.

Table VIII-E-2

Degrees of Excellence: First Set

Phrase	Scale Value	SD
Superior	20.12	1.17
Fantastic	20.12	0.83
Tremendous	19.84	1.31
Superb	19.80	1.19
Excellent	19.40	1.73
Terrific	19.00	2.45
Outstanding	18.96	1.99
Wonderful	17.32	2.30
Delightful	16.92	1.85
Fine	14.80	2.12
Good	14.32	2.08
Pleasant	13.44	2.06
Nice	12.56	2.14
Acceptable	11.12	2.59
Average	10.84	1.55
All right	10.76	1.42
O.K.	10.28	1.67
Neutral	9.80	1.50
Fair	9.52	2.06
Mediocre	9.44	1.80
Unpleasant	5.04	2.82
Bad	3.88	2.19
Very bad	3.20	2.10
Unacceptable	2.64	2.04
Awful	1.92	1.50
Terrible	1.76	.77
Horrible	1.48	.87

From: Myers and Warner (1968).
See Section VIII-A 17.

Table VIII-E-3

Degrees of Excellence: Second Set

Phrase	Scale Value	SD
Best of all	6.15	2.48
Excellent	3.71	1.01
Wonderful	3.51	.97
Mighty fine	2.88	.67
Especially good	2.86	.82
Very good	2.56	.87
Good	1.91	.76
Pleasing	1.58	.65
O.K.	.87	1.24
Fair	.78	.85
Only fair	.71	.64
Not pleasing	-.83	.67
Poor	-1.55	.87
Bad	-2.02	.80
Very bad	-2.53	.64
Terrible	-3.09	.98

From: Jones and Thurstone (1955).
See Section VIII-A 12.

Table VIII-E-4

Degrees of Like and Dislike

Phrase	Scale Value	SD
Like extremely	4.16	1.62
Like intensely	4.05	1.59
Strongly like	2.96	.69
Like very much	2.91	.60
Like very well	2.60	.78
Like quite a bit	2.32	.52
Like fairly well	1.51	.59
Like	1.35	.77
Like moderately	1.12	.61
Mildly like	.85	.47
Like slightly	.69	.32
Neutral	.02	.18
Like not so well	-.30	1.07
Like not so much	-.41	.94
Dislike slightly	-.59	.27
Mildly dislike	-.74	.35
Dislike moderately	-1.20	.41
Dislike	-1.58	.94
Don't like	-1.81	.97
Strongly dislike	-2.37	.53
Dislike very much	-2.49	.64
Dislike intensely	-3.33	1.39
Dislike extremely	-4.32	1.86

From: Jones and Thurstone (1955).
See Section VIII-A 12.

Table VIII-E-5

Degrees of Good and Poor

Phrase	Scale Value	SD
Exceptionally good	18.56	2.36
Extremely good	18.44	1.61
Unusually good	17.08	2.43
Remarkably good	16.68	2.19
Very good	15.44	2.77
Quite good	14.44	2.76
Good	14.32	2.08
Moderately good	13.44	2.23
Reasonably good	12.92	2.93
Fairly good	11.96	2.42
Slightly good	11.84	2.19
So-so	10.08	1.87
Not very good	6.72	2.82
Moderately poor	6.44	1.64
Reasonably poor	6.32	2.46
Slightly poor	5.92	1.96
Poor	5.72	2.09
Fairly poor	5.64	1.68
Quite poor	4.80	1.44
Unusually poor	3.20	1.44
Very poor	3.12	1.17
Remarkably poor	2.88	1.74
Exceptionally poor	2.52	1.19
Extremely poor	2.08	1.19

From: Myers and Warner (1968).
See Section VIII-A 17.

Table VIII-E-6

Degrees of Good and Bad

Phrase	Scale Value
Extremely good	3.449
Very good	3.250
Unusually good	3.243
Decidedly good	3.024
Quite good	2.880
Rather good	2.755
Good	2.712
Pretty good	2.622
Somewhat good	2.462
Slightly good	2.417
Slightly bad	1.497
Somewhat bad	1.323
Rather bad	1.232
Bad	1.024
Pretty bad	1.018
Quite bad	.924
Decidedly bad	.797
Unusually bad	.662
Very bad	.639
Extremely bad	.470

From: Cliff (1959).
See Section VIII-A 6.

Table VIII-E-7

Degrees of Agree and Disagree

Phrase	Mean	SD
Decidedly agree	2.77	.41
Quite agree	2.37	.49
Considerably agree	2.21	.42
Substantially agree	2.10	.50
Moderately agree	1.47	.41
Somewhat agree	.94	.41
Slightly agree	.67	.36
Perhaps agree	.52	.46
Perhaps disagree	-.43	.46
Slightly disagree	-.64	.38
Somewhat disagree	-.93	.47
Moderately disagree	-1.35	.42
Quite disagree	-2.16	.57
Substantially disagree	-2.17	.51
Considerably disagree	-2.17	.45
Decidedly disagree	-2.76	.43

From: Altemeyer (1970).
See Section VIII-A 1.

Table VIII-E-8

Degrees of More and Less

Phrase	Scale Value	Inter-quartile Range
Very much more	8.02	.61
Much more	7.67	1.04
A lot more	7.50	1.06
A good deal more	7.29	.98
More	6.33	1.01
Somewhat more	6.25	.98
A little more	6.00	.58
Slightly more	5.99	.57
Slightly less	3.97	.56
A little less	3.96	.54
Less	3.64	1.04
Much less	2.55	1.06
A good deal less	2.44	1.11
A lot less	2.36	1.03
Very much less	1.96	.52

From: Dodd and Gerberick (1960).
See Section VIII-A 7.

Table VIII-E-9

Degrees of Adequate and Inadequate

Phrase	Mean	SD
Totally adequate	4.620	.846
Absolutely adequate	4.540	.921
Completely adequate	4.490	.825
Extremely adequate	4.412	.719
Exceptionally adequate	4.380	.869
Entirely adequate	4.340	.863
Wholly adequate	4.314	1.038
Fully adequate	4.294	.914
Very very adequate	4.063	.876
Perfectly adequate	3.922	1.026
Highly adequate	3.843	.606
Most adequate	3.843	.978
Very adequate	3.420	.851
Decidedly adequate	3.140	1.536
Considerably adequate	3.020	.874
Quite adequate	2.980	.979
Largely adequate	2.863	.991
Substantially adequate	2.608	1.030
Reasonably adequate	2.412	.771
Pretty adequate	2.306	.862
Rather adequate	1.755	.893
Mildly adequate	1.571	.670
Somewhat adequate	1.327	.793
Slightly adequate	1.200	.566
Barely adequate	.627	.928
Neutral	.000	.000
Borderline	-.020	.316
Barely inadequate	-1.157	.638
Mildly inadequate	-1.353	.621
Slightly inadequate	-1.380	.772
Somewhat inadequate	-1.882	.732
Rather inadequate	-2.102	.974
Moderately inadequate	-2.157	1.017
Fairly inadequate	-2.216	.800
Pretty inadequate	-2.347	.959
Considerably inadequate	-3.600	.680
Very inadequate	-3.735	.777
Decidedly inadequate	-3.780	.944
Most inadequate	-3.980	1.545
Highly inadequate	-4.196	.741

(Table continued on next page)

Table VIII-E-9 (Cont.)

Degrees of Adequate and Inadequate

Phrase	Mean	SD
Very very inadequate	-4.460	.537
Extremely inadequate	-4.608	.527
Fully inadequate	-4.667	.676
Exceptionally inadequate	-4.680	.508
Wholly inadequate	-4.784	.498
Entirely inadequate	-4.792	.644
Completely inadequate	-4.800	.529
Absolutely inadequate	-4.880	.431
Totally inadequate	-4.900	.412

From: Matthews, Wright, and Yudowitch (1975).
See Section VIII-A 13.

Table VIII-E-10

Degrees of Acceptable and Unacceptable

Phrase	Mean	SD
Wholly acceptable	4.725	.563
Completely acceptable	4.686	.610
Fully acceptable	4.412	.867
Extremely acceptable	4.392	.716
Most acceptable	4.157	.915
Very very acceptable	4.157	.825
Highly acceptable	4.040	.631
Quite acceptable	3.216	.956
Largely acceptable	3.137	.991
Acceptable	2.392	1.456
Reasonably acceptable	2.294	.722
Moderately acceptable	2.280	.722
Pretty acceptable	2.000	1.125

(Table continued on next page)

Table VIII-E-10 (Cont.)

Degrees of Acceptable and Unacceptable

Phrase	Mean	SD
Rather acceptable	1.939	.818
Fairly acceptable	1.840	.924
Mildly acceptable	1.686	.700
Somewhat acceptable	1.458	1.241
Barely acceptable	1.078	.518
Slightly acceptable	1.039	.522
Sort of acceptable	.940	.645
Borderline	.000	.200
Neutral	.000	.000
Marginal	-.120	.515
Barely unacceptable	-1.100	.300
Slightly unacceptable	-1.255	.589
Somewhat unacceptable	-1.765	.674
Rather unacceptable	-2.020	.836
Fairly unacceptable	-2.160	.880
Moderately unacceptable	-2.340	.681
Pretty unacceptable	-2.412	.662
Reasonably unacceptable	-2.440	.753
Unacceptable	-2.667	1.381
Substantially unacceptable	-3.235	.899
Quite unacceptable	-3.388	1.066
Largely unacceptable	-3.392	.818
Considerably unacceptable	-3.440	.779
Notably unacceptable	-3.500	1.044
Decidedly unacceptable	-3.837	1.017
Highly unacceptable	-4.294	.535
Most unacceptable	-4.420	.724
Very very unacceptable	-4.490	.500
Exceptionally unacceptable	-4.540	.607
Extremely unacceptable	-4.686	.464
Completely unacceptable	-4.900	.361
Entirely unacceptable	-4.900	.361
Wholly unacceptable	-4.922	.269
Absolutely unacceptable	-4.922	.334
Totally unacceptable	-4.941	.235

From: Matthews, Wright, and Yudowitch (1975).
See Section VIII-A 13.

Table VIII-E-11

Comparison Phrases

Phrase	Mean	SD
Best of all	4.896	.510
Absolutely best	4.843	.459
Truly best	4.600	.721
Undoubtedly best	4.569	.823
Decidedly best	4.373	.839
Best	4.216	1.459
Absolutely better	4.060	.988
Extremely better	3.922	.882
Substantially best	3.700	.922
Decidedly better	3.412	.933
Conspicuously better	3.059	.802
Moderately better	2.255	.737
Somewhat better	1.843	.801
Rather better	1.816	.719
Slightly better	1.157	.776
Barely better	.961	.656
Absolutely alike	.588	1.623
Alike	.216	.847
The same	.157	.801
Neutral	.000	.000
Borderline	-.061	.314
Marginal	-.184	.919
Barely worse	-1.039	.816
Slightly worse	-1.216	.498
Somewhat worse	-2.078	.860
Moderately worse	-2.220	.944
Noticeably worse	-2.529	1.036
Worse	-2.667	1.423
Notably worse	-3.020	1.038
Largely worse	-3.216	1.108
Considerably worse	-3.275	1.206
Conspicuously worse	-3.275	.887
Much worse	-3.286	.808
Substantially worse	-3.460	.899
Decidedly worse	-3.760	.907
Very much worse	-3.941	.752
Absolutely worse	-4.431	.823
Decidedly worst	-4.431	.748
Undoubtedly worst	-4.510	.872
Absolutely worst	-4.686	1.291
Worst of all	-4.776	1.298

From: Matthews, Wright, and Yudowitch (1975).
See Section VIII-A 13.

Table VIII-E-12

Degrees of Satisfactory and Unsatisfactory

Phrase	Scale Value	SD
Quite satisfactory	4.35	.95
Satisfactory	3.69	.87
Not very satisfactory	2.11	.76
Unsatisfactory but usable	2.00	.87
Very unsatisfactory	.69	1.32

From: U.S. Army (1973). See Section VIII-A 21.

Table VIII-E-13

Degrees of Unsatisfactory

Phrase	Scale Value
Unsatisfactory	1.47
Quite unsatisfactory	1.00
Very unsatisfactory	.75
Unusually unsatisfactory	.75
Highly unsatisfactory	.71
Very, very unsatisfactory	.25
Extremely unsatisfactory	.10
Completely unsatisfactory	.00

From: Mosier (1941).
See Section VIII-A 16.

Table VIII-E-14

Degrees of Pleasant

Phrase	Scale Value
Extremely pleasant	3.490
Very pleasant	3.174
Unusually pleasant	3.107
Decidedly pleasant	3.028
Quite pleasant	2.849
Pleasant	2.770
Rather pleasant	2.743
Pretty pleasant	2.738
Somewhat pleasant	2.505
Slightly pleasant	2.440

From: Cliff (1959).
See Section VIII-A 6.

Table VIII-E-15

Degrees of Agreeable

Phrase	Scale Value
Very, very agreeable	5.34
Extremely agreeable	5.10
Highly agreeable	5.02
Completely agreeable	4.96
Unusually agreeable	4.86
Very agreeable	4.82
Quite agreeable	4.45
Agreeable	4.19

From: Mosier (1941).
See Section VIII-A 16.

Table VIII-E-16

Degrees of Desirable

Phrase	Scale Value
Very, very desirable	5.66
Extremely desirable	5.42
Completely desirable	5.38
Unusually desirable	5.23
Highly desirable	5.15
Very desirable	4.96
Quite desirable	4.76
Desirable	4.50

From: Mosier (1941).
See Section VIII-A 16.

Table VIII-E-17

Degrees of Nice

Phrase	Scale Value
Extremely nice	3.351
Unusually nice	3.155
Very nice	3.016
Decidedly nice	2.969
Pretty nice	2.767
Quite nice	2.738
Nice	2.636
Rather nice	2.568
Somewhat nice	2.488
Slightly nice	2.286

From: Cliff (1959).
See Section VIII-A 6.

Table VIII-E-18

Degrees of Adequate

Phrase	Scale Value	SD
More than adequate	4.13	1.11
Adequate	3.39	.87
Not quite adequate	2.40	.85
Barely adequate	2.10	.84
Not adequate	1.83	.98

From: U.S. Army (1973).
See Section VIII-A 21.

Table VIII-E-19

Degrees of Ordinary

Phrase	Scale Value
Ordinary	2.074
Very ordinary	2.073
Somewhat ordinary	2.038
Rather ordinary	2.034
Pretty ordinary	2.026
Slightly ordinary	1.980
Decidedly ordinary	1.949
Extremely ordinary	1.936
Unusually ordinary	1.875

From: Cliff (1959).
See Section VIII-A 6.

Table VIII-E-20

Degrees of Average

Phrase	Scale Value
Rather average	2.172
Average	2.145
Quite average	2.101
Pretty average	2.094
Somewhat average	2.080
Unusually average	2.062
Extremely average	2.052
Very average	2.039
Slightly average	2.023
Decidedly average	2.020

From: Cliff (1959).
See Section VIII-A 6.

Table VIII-E-21

Degrees of Hesitation

Phrase	Scale Value	Inter-quartile Range
Without hesitation	7.50	6.54
With little hesitation	5.83	3.40
Hesitant	4.77	1.06
With some hesitation	4.38	1.60
With considerable hesitation	3.29	3.39
With much hesitation	3.20	5.25
With great hesitation	2.41	6.00

From: Dodd and Gerberick (1960). See Section VIII-A 7.

Table VIII-E-22

Degrees of Inferior

Phrase	Scale Value
Slightly inferior	1.520
Somewhat inferior	1.516
Inferior	1.323
Rather inferior	1.295
Pretty inferior	1.180
Quite inferior	1.127
Decidedly inferior	1.013
Unusually inferior	.963
Very inferior	.927
Extremely inferior	.705

From: Cliff (1959).
See Section VIII-A 6.

Table VIII-E-23

Degrees of Poor

Phrase	Scale Value
Poor	1.60
Quite poor	1.30
Very poor	1.18
Unusually poor	.95
Extremely poor	.95
Completely poor	.92
Very, very poor	.55

From: Mosier (1941).
See Section VIII-A 16.

Table VIII-E-24

Descriptive Phrases

Phrase	Scale Value	Inter-quartile Range
Complete	8.85	.65
Extremely vital	8.79	.84
Very certain	8.55	1.05
Very strongly	8.40	1.04
Very crucial	8.29	1.12
Very important	8.22	1.16
Very sure	8.15	.95
Almost complete	8.06	.58
Of great importance	8.05	.91
Very urgent	8.00	.90
Feel strongly toward	7.80	1.60
Essential	7.58	1.85
Very vital	7.55	1.05
Certain	7.13	1.44
Strongly	7.07	.67
Important	6.83	1.14
Good	6.72	1.20
Urgent	6.41	1.53
Crucial	6.39	1.73
Sure	5.93	1.87
Vital	5.92	1.63
Moderately	5.24	.99
Now	5.03	.53
As at present	5.00	.50
Fair	4.96	.77
Don't know	4.82	.82
Undecided	4.76	1.06
Don't care	4.63	2.00
Somewhat	3.79	.94
Indifferent	3.70	2.20
Object strongly to	3.50	6.07
Not important	3.09	1.33
Unimportant	1.94	1.42
Bad	2.83	.93
Uncertain	2.83	2.50
Doesn't make any difference	2.83	3.13
Not sure	2.82	1.24
Not certain	2.64	2.62

(Table continued on next page)

Table VIII-E-24 (Cont.)

Descriptive Phrases

Phrase	Scale Value	Inter-quartile Range
Non-essential	2.58	1.67
Doesn't mean anything	2.50	2.71
Insignificant	2.12	1.14
Very little	2.08	.64
Almost none	2.04	.57
Very unimportant	1.75	1.25
Only as a last resort	1.70	7.30
Very bad	1.50	1.13
None	1.11	.59

From: Dodd and Gerberick (1960).
See Section VIII-A 7.

F. Sample Sets of Response Alternatives

It is sometimes valuable and is a time-saver to have lists of response alternatives available to use. The tables in this section give some examples of response alternatives that have been selected on different bases. These sets do not exhaust all possibilities.

The sets of response alternatives that appear in Table VIII-F-1 were selected so that the phrases in each set would have means at least one standard deviation away from each other and have parallel wording. Some of the sets of response alternatives have extreme endpoints, some do not. The sets of response alternatives shown in Table VIII-F-2 were selected so that the phrases in each set would be as nearly equally distant from each other as possible without regard to parallel wording. Table VIII-F-3 contains sets of response alternatives selected from lists of descriptors with only scale values given. The phrases were selected on the bases of equal appearing intervals. Table VIII-F-4 has sets of response alternatives selected from order of merit lists of descriptors.

Table VIII-F-1

Sets of Response Alternatives Selected So Phrases Are at Least One Standard Deviation Apart and Have Parallel Wording

Set No.	Response Alternatives
1.	Completely acceptable Reasonably acceptable Barely acceptable Borderline Barely unacceptable Reasonably unacceptable Completely unacceptable
2.	Wholly acceptable Largely acceptable Borderline Largely unacceptable Wholly unacceptable
3.	Largely acceptable Barely acceptable Borderline Barely unacceptable Largely unacceptable
4.	Reasonably acceptable Slightly acceptable Borderline Slightly unacceptable Reasonably unacceptable
5.	Totally adequate Very adequate Barely adequate Borderline Barely inadequate Very inadequate Totally inadequate
6.	Completely adequate Considerably adequate Borderline Considerably inadequate Completely inadequate
7.	Very adequate Slightly adequate Borderline Slightly inadequate Very inadequate
8.	Highly adequate Mildly adequate Borderline Mildly inadequate Highly inadequate
9.	Decidedly agree Substantially agree Slightly agree Slightly disagree Substantially disagree Decidedly disagree
10.	Moderately agree Perhaps agree Neutral Perhaps disagree Moderately disagree
11.	Undoubtedly best Conspicuously better Moderately better Alike Moderately worse Conspicuously worse Undoubtedly worst
12..	Moderately better Barely better The same Barely worse Moderately worse

(Table continued on next page)

Table VIII-F-1 (Cont.)

Sets of Response Alternatives Selected So Phrases Are at Least One Standard Deviation Apart and Have Parallel Wording

Set No.	Response Alternatives	Set No.	Response Alternatives
13.	Extremely good	16.	Like extremely
	Remarkably good		Like moderately
	Good		Neutral
	So-so		Dislike moderately
	Poor		Dislike extremely
	Remarkably poor		
	Extremely poor	17.	Strongly like
			Like
14.	Exceptionally good		Neutral
	Reasonably good		Don't like
	So-so		Strongly dislike
	Reasonably poor		
	Exceptionally poor	18.	Very much more
			A good deal more
15.	Very important		A little more
	Important		A little less
	Not important		A good deal less
	Very unimportant		Very much less

Table VIII-F-2

Sets of Response Alternatives Selected So That
Intervals Between Phrases Are as Nearly Equal as Possible

Set No.	Response Alternatives	Set No.	Response Alternatives
1.	Completely acceptable Reasonably acceptable Borderline Moderately unacceptable Extremely unacceptable	7.	Perfect in every respect Very good Good Could use some minor changes Not very good Better than nothing Extremely poor
2.	Totally adequate Pretty adequate Borderline Pretty inadequate Extremely inadequate	8.	Excellent Good Only fair Poor Terrible
3.	Highly adequate Rather adequate Borderline Somewhat inadequate Decidedly inadequate	9.	Extremely good Quite good So-so Slightly poor Extremely poor
4.	Quite agree Moderately agree Perhaps agree Perhaps disagree Moderately disagree Substantially disagree	10.	Remarkably good Moderately good So-so Not very good Unusually poor
5.	Undoubtedly best Moderately better Borderline Noticeably worse Undoubtedly worst	11.	Without hesitation With little hesitation With some hesitation With great hesitation
6.	Fantastic Delightful Nice Mediocre Unpleasant Horrible	12.	Strongly like Like quite a bit Like Neutral Mildly dislike Dislike very much Dislike extremely

(Table continued on next page)

Table VIII-F-2 (Cont.)

Sets of Response Alternatives Selected So That
Intervals Between Phrases Are as Nearly Equal as Possible

Set No.	Response Alternatives	Set No.	Response Alternatives
13.	Like quite a bit Like Like slightly Borderline Dislike slightly Dislike moderately Don't like	15.	Very much more A little more Slightly less Very much less
14.	Like quite a bit Like fairly well Borderline Dislike moderately Dislike very much		

Table VIII-F-3

Sets of Response Alternatives Selected
from Lists Giving Scale Values Only

Set No.	Response Alternatives	Set No.	Response Alternatives
1.	Very, very agreeable Usually agreeable Quite agreeable Agreeable	6.	Extremely nice Decidedly nice Nice Slightly nice
2.	Rather average Quite average Unusually average Decidedly average	7.	Ordinary Slightly ordinary Unusually ordinary
3.	Very, very desirable Completely desirable Very desirable Desirable	8.	Extremely pleasant Decidedly pleasant Somewhat pleasant
4.	Extremely good Somewhat good Slightly bad Extremely bad	9.	Poor Very poor Very, very poor
5.	Slightly inferior Rather inferior Unusually inferior Extremely inferior	10.	Very, very agreeable Extremely agreeable Very agreeable Quite agreeable Agreeable

Note. Selected so that intervals between phrases are as equal as possible.

Table VIII-F-4

Sets of Response Alternatives Selected
Using Order of Merit Lists of Descriptor Terms

Set No.	Response Alternatives
1.	Very good
	Good
	Borderline
	Poor
	Very poor
2.	Very satisfactory
	Satisfactory
	Borderline
	Unsatisfactory
	Very unsatisfactory
3.	Very superior
	Superior
	Borderline
	Poor
	Very poor
4.	Extremely useful
	Of considerable use
	Of use
	Not very useful
	Of no use

Chapter IX: Physical Characteristics of Questionnaires

A. Overview

This chapter considers five topics related to the physical characteristics of questionnaires: the location of response alternatives relative to the stem (Section IX-B); questionnaire length (Section IX-C); questionnaire format considerations (Section IX-D); the use of answer sheets (Section IX-E); and the use of branching (Section IX-E).

B. Location of Response Alternatives Relative to the Stem

Research to determine what effect the location of response alternatives relative to the question stem has on subjects' responses is practically nonexistent. There is some evidence, however, that untrained raters can make relatively error-free graphic ratings regardless of whether the "good" end of the scale is at the left, right, top, or bottom.

In designing a specific questionnaire, the following points should be considered regarding the location of response alternatives relative to the stem:

1. With multiple choice items, the response alternatives are usually arranged vertically under the stem as shown in Figure IV-C-1. With a large number of response alternatives, two or more columns of vertically arranged alternatives might be used. Sometimes, if there are only two or three alternatives (such as "Yes" and "No"), they are placed horizontally rather than vertically.

2. Graphic rating scales are usually placed horizontally on a page. However, the descriptive words, phrases, or sentences on a scale should be concentrated as much as possible at specific points on the scale. This is usually easier if the scales are placed vertically on the page, but it can be done either way. Descriptors need not be equally spaced along graphic scales, and should not be if there is reason to believe the psychological distances between them are not equal.

3. With nongraphic (or "numerical") rating scale items and with ranking and forced choice items, the response alternatives are usually placed vertically under the question stem. See examples in Chapter IV. Sometimes rating scale items are placed horizontally under the stem as shown in Figure VII-B-1. If a number of rating scale items all use the same response alternatives, the question stems can be presented in a column with the response alternatives to the right as shown in Figure IX-B-1.

   In Figure IX-B-1, the response alternatives have been rotated 90 degrees to save space. An effort should be made to place the response alternative horizontal with the bottom of the page so that the respondent does not need to turn the page sideways to read them.

4. The response alternatives for semantic differential items are usually placed horizontally on the page. For an example, see Figure IV-I-1.

5. Use precoded cards with alphabet letters for responses to items with sensitive content that might be viewed as threatening. This would be appropriate for questionnaires administered by personal interview. Selection by respondent of the alphabet letter can later be transposed to numbers for analysis purposes. This technique is used to obtain less distortion to reduce the social desirability response set.

Figure IX-B-1

Arrangement of Items With Same
Rating Scale Response Alternatives

1. How satisfied or dissatisfied are you with each of the following factors or things?

	Very Satisfied	Satisfied	Borderline	Dissatisfied	Very Dissatisfied
a. Type of furniture in barracks.	___	___	___	___	___
b. Medical service to soldiers.	___	___	___	___	___
c. Quality of mess hall food.	___	___	___	___	___
d. Leadership of generals.	___	___	___	___	___
e. Opportunity for promotion.	___	___	___	___	___
f. Army pay.	___	___	___	___	___
g. Civilian opinion of Army.	___	___	___	___	___

6. For respondents with a low education level, an easy format with stems and anchors easy to understand is essential. Sometimes, respondents have a preference for questionnaire formats with which they have had previous experience. However, preference for specific kinds of formats does not mean that the results will be more reliable. In some studies, respondents were more accurate in their ratings with less preferred formats.

C. Questionnaire Length

1. General

The length of questionnaires used in field tests has ranged from one page to as many as 30 pages, perhaps more. How long can one expect a respondent to work effectively at the questionnaire-answering task? At what point do attention and motivation start to degrade, thereby producing poorly considered responses or the omission of responses? Research information on this point is not available to provide a basis for a firm recommendation. There is even disagreement on the effect of questionnaire length on the response rate to mailed questionnaires. The number of items and number of pages in a questionnaire may not necessarily be related to response rate for mailed questionnaires.

However, questionnaires which require longer than one hour to complete will, in most situations, cause boredom and indifference. Even 10 or 15 minutes may be too long if the questionnaire is perceived by the respondent as redundant or asking unnecessary questions. If one is concerned about the effects of a long questionnaire, alternate forms should be used, wherein the order of items is reversed (or approximately so). For example, the items answered last on 50% of the forms would be answered first on the other 50% of the forms. One could also split the respondent group in half and give half of the questions to each group--provided that the two groups were fairly equivalent in relevant characteristics. Splitting the respondent group in half increases the complexity of the survey and may affect the precision of the measures. It is assumed that everything else would already have been done to reduce the number of items before one of these approaches is used.

For questionnaires administered by interview, survey guidelines were established by the Federal Office of Management and Budget. They suggest that interviews should not take longer than half an hour, although there may be valid reasons why more information would be required. This would, of course, extend the interview length. Many surveys take an hour or longer to complete. There are no firm guidelines. Pretesting the questionnaire may provide data on the effect of number of items on the response rate.

2. Results of a Study

In a 1976 study, ARI assisted TCATA in obtaining and analyzing questionnaire responses from a group of trainees whose duration and location of basic and advanced individual training was handled differently from the usual. The number of trainees answering items 1-7 and 48-54 of a 54-item questionnaire is shown below. Note that there is very little drop in the number of men in either group as we skip from items 1-7 to items 48-52. This suggests that a 50-item questionnaire, administered as this was, was not so long that persons stopped responding after answering successively more questions.

Now note the sharp drop-about 15% and 9% for the two groups-in responses to items 53 and 54. A more gradual decrease in number of people responding is more what one would expect if they are being "worn down" or fatigued by excessive length.

This result was puzzling, but then it was noted that items 53 and 54 are alone together on the tenth and final page of the questionnaire. It is speculated that many/most of those not answering items 53 and 54 turned page 10 over along with page 9 and thought they had answered all that was required of them. No one checked their questionnaires when they were handed in to see if they had left any items blank. The reductions in respondents appears more of a "last page phenomena" than a consequence of an excessively long questionnaire.

Item #	Experimental Group	Control Group
1	716	512
2	716	513
3	717	511
4	714	513
5	716	514
6	713	510
7	716	511
:	:	:
48	707	509
49	707	508
50	707	508
51	707	510
52	698	505
53	593	462
54	604	461

D. Questionnaire Format Considerations

This section addresses the format of questionnaire items, title and other identification marks, printed introductions, planning to facilitate processing, and other questionnaire format considerations.

1. Format of Questionnaire Items and Format Bias

Item format biases occur when responses to items (questions) are influenced by the question stem or response alternatives. The following guidance is provided:

a. The format of all questionnaire items on a questionnaire should be consistent whenever possible. Mixing multiple choice questions, open-ended questions, scales, etc. is normally not desirable.

b. Punctuation and question structure should be consistent and in accordance with proper sentence structure principles. Where incomplete sentences (e.g., "The training that I have received at Fort Hood has been" with five response alternatives of "very challenging" through "very unchallenging") are used as stems, no extraneous punctuation, such as a colon, need be put at the end of the stem. The first word of the response alternatives should not be capitalized unless they would be if the statement were written as a continuous sentence. Terminal punctuation at the end of the response alternatives should follow the same general rule of consistency with normal sentence structure. Hence, a period would ordinarily be placed after each response alternative.

When an item consists of a complete question (e.g., "How satisfied or dissatisfied are you with the furniture in the barracks?"), the first word of the response alternatives should be capitalized since it does not continue a sentence. If the response alternatives constitute complete sentences, then they should have periods at the end, or whatever other terminal punctuation is appropriate. Sometimes periods are placed at the end of extremely long response alternatives even if they are not sentences. Ordinarily, then, with this form of item, periods would not be placed after the response alternatives.

Exceptions to the above suggestions should be made whenever the exception would improve clarity. An example might be when periods would be confused with decimal points.

c. When items are ambiguous, a recognizable pattern of inappropriate responses is often produced.

d. Item format bias may be a function of how items are sequenced and grouped.

e. Some authors conclude that a bias can be expected from all closed-end questions where answers must be selected from two or more fixed choices.

f. The paired-comparison format may be useful for those respondents who tend to check many items from a list, and for those who check only a few.

g. Card sorting may show the least item format bias.

h. With two-way choices, some respondents have a tendency to select the first alternative. Others have a tendency to select the second. With other multiple choice items, some respondents have a tendency to select certain categories.

i. There is some evidence that the first response alternative to a question is chosen somewhat more frequently than the others.

j. Two studies were conducted by Mayer and Piper (1982) regarding physical layout of a questionnaire. Questionnaire layout can be confusing to respondents. The wrong categories were initially marked by respondents indicating erroneous brand preferences. An example is provided in Figure IX-D-1 to illustrate how modification of the questionnaire format facilitated clarification. The questionnaire layout that confused respondents did not have a response alternative for "Other brand." The layout was identical to that of Brand A through Brand G response alternatives. There was no bracketed response alternative for "Other brand."

Mayer, C. S., & Piper, C. (1982). A note on the importance of layout in self-administered questionnaires. Journal of Marketing Research, 19(3), 390-391.

Figure IX-D-1

Original Questionnaire Format

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

Modified Questionnaire Format

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

2. Title and Other Identification Marks

Each questionnaire should carry a descriptive title centered at the top of the first page of questions and on the instructional and/or introductory cover page if such is used. Each questionnaire form should also be designated by form number to distinguish it from other forms. This number usually goes in the upper left-hand corner of each page.

3. Printed Introductions

Introductions are sometimes printed at the start of a questionnaire to tell respondents the purpose and importance of the questionnaire, and the importance of their cooperation in answering all questions carefully. Methodological research is needed to determine the effectiveness of such introductions, but if they are too lengthy, there is always the possibility that they might be counterproductive. Regardless, if the introduction is going to run more than a quarter of a page, it might better be placed on a cover sheet.

See Section X-B regarding the content of questionnaire instructions.

4. Planning to Facilitate Processing

Where possible, questionnaires should be designed to facilitate data collection, reduction, and analysis. This frequently involves formulating the questionnaire for machine processing. For small samples, however, manual processing should normally be employed since the effort needed to plan for machine processing is not justified by anticipated data reduction time savings. How to format a questionnaire for machine processing is outside the current scope of this manual. See Section IX-E regarding the use of answer sheets.

5. Other Questionnaire Format Considerations

a. If the respondent's name, rank, etc. is really needed, ask for it on the front page. (See also Section X-C.) Sometimes other information is needed about respondents so that it can be correlated with their responses. This may include duty MOS, special army training, combat experience, etc. If it is really needed, it is usually asked for on the front page along with name.

b. If a questionnaire has over two pages, page numbers should be used. They are ordinarily put at the center bottom of each page.

c. A questionnaire should not be crowded or cluttered in appearance. If it is, certain items might be missed.

d. Each item in a questionnaire should be numbered or lettered so it can be identified and referred to.

e. Sufficient room should be left for the respondent to write in answers to open-ended questions.

f. Directions should be well displayed and unmistakably clear.

g. There is research evidence that an attractive questionnaire increases response rates.

h. Different colored pages or questionnaire forms may aid in the sorting of data and may have appeal to the respondents.

E. Use of Answer Sheets

As noted in Section IX-D 4, when possible, questionnaires should be designed to facilitate data collection, processing, and analyses. Hence, if the number of questions warrant it, consideration should be given to the use of separate answer sheets. An answer sheet can be designed for either hand or machine processing.

When considering the possible use of answer sheets, the following points should be kept in mind:

1. The use of a separate answer sheet may require additional or different abilities than responding on the questionnaire itself.

2. Depending upon their prior experiences with them, respondents may find it more difficult to use a separate answer sheet than to respond on the questionnaire sheet.

3. It is normally more difficult and time-consuming for the respondent to use a separate answer sheet. However, separate answer sheets have been used successfully for some purposes.

4. When separate answer sheets are employed, the questionnaire booklets are reusable.

5. Respondents sometimes err in using the last spaces on a multiple choice answer sheet when there are more spaces than response alternatives. This can be avoided by the use of tailor-made sheets.

F. Use of Branching

Some questionnaires are constructed so that respondents need not answer every question in the survey. Branching is used to guide respondents through a survey instrument to appropriate questions. This technique requires the construction of questions which are integrated and then arranged to implement the purposes of the branching.

1. When to Use Branching

Branching is used when the researcher wants to screen respondents and assign them to subgroups. It is also a way to guard against having the respondents' answers be influenced by the question(s) that are bypassed in the branching (or branched around). This is known as position effect. The questionnaire educates the respondent on the topic which it covers. Branching can be used to measure the effects of the questionnaire educating the respondent. The more forward branching that occurs, the less education is being given to the respondents (they don't need it).

Branching can be used to reduce interview time for questionnaires administered by interview. Clear branching instructions are required for the interviewer, as well as interviewer training. Self-administered/group-administered questionnaires which use branching can reduce the time to complete for the respondents. However, there is a greater risk of item nonresponse following a branch for these questionnaires.

2. Filter Questions

Filter questions are developed to determine how respondents are to be guided through the questionnaire. Filter questions screen out respondents from certain sets of questions on the questionnaire. Branching is used so that the respondent can be routed into another subset of questions. Consequently, the survey functions as a set of filters through which some respondents pass while others are detained or routed into different topic area(s).

3. Branching Applications

Respondents are routed through the questionnaire by presenting them with more difficult or more concrete questions. This approach forces the respondent to consider the topic area from many viewpoints. Clear branching instructions are required for all questionnaires. Items immediately following a branch tend to have an increased rate of nonresponse; to this extent, the branching instructions were unclear or not understood.

4. Recommendations

Surveys conducted by interview which use branching may be choppy. Interviewers require smooth transitions between branches, and training in conducting the survey. Branching is best used to reduce interview time. Branching may be used to reduce/avoid exposing the respondent to items that are irrelevant or non-essential. This forces the interviewer to ask only pertinent questions regardless of the interviewer's persuasion. Branching for mail surveys and group-administered surveys has a greater probability of increasing item nonresponse rate than a survey conducted by interview.

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

When branching questionnaires are used to measure respondent attitudes, there is a greater possibility for introducing bias into the data. Questionnaires covering topic areas dealing with fact instead of attitude are preferred when branching is used. Branching may be used to reduce/avoid exposing the respondent to items that are irrelevant or nonessential. This forces the interviewer to ask only pertinent questions regardless of the interviewer's persuasion. See Sections VI-F-1 and X-D-3.

Chapter X: Considerations Related to Questionnaire Administration

A. Overview

Considerations related to the administration of questionnaires are discussed in this chapter. Such matters are obviously of concern when questionnaires are constructed. Questionnaire instructions are discussed in Section X-B, anonymity for respondents in Section X-C, and motivational factors related to questionnaire administration in Section X-D. Administration time, characteristics of administrators, and administrative conditions are the topics of Section X-E, X-F, and X-G, respectively. The training of raters and other evaluators is the concern of Section X-H, while other factors related to questionnaire administration are considered in Section X-I.

B. Instructions

Care must be exercised in preparing instructions for questionnaires since they are quite likely to affect the way the respondent answers the questions. For example, even mildly anger-arousing printed instructions may elicit responses of negativism.

Although further research is needed to fully determine the influence of instructions on responses, some practical guidelines can be offered:

1. It is sometimes preferred that an oral statement of questionnaire purpose be given to respondents. If this is not practical or a person with appropriate credibility and/or status cannot be supplied to make the statement, then a printed statement must suffice. (See Section IX-D 3 regarding printed introductions.)

2. Lengthy instructions for completing questionnaires should be avoided. They may tend to confuse the respondents rather than help them.

3. The option of orally presenting instructions is often available. When oral instructions are given, they are usually given just prior to administering the questionnaire.

4. If instructions are given orally and an illustration is needed, a visual display should be available which may include a printed version of more complex instructions.

5. When questionnaires are group-administered, it should be announced that aides will check each respondent's questionnaire for completeness, if such a process can be implemented.

6. "Cute" examples on instructions should not be used. They will damage rapport and detract from the seriousness of the questionnaires, particularly for more mature and older respondents. It is best to use a neutral example that will be suitable for all respondents.

7. Obviously, instructions should be given in a way that all respondents can understand them. Care should be exercised about the level of vocabulary used.

An example is given on the following page of the instructions that might precede the items of a questionnaire. In this example, the responses were to be given on a separate "answer" or response sheet.

TRAINING ATTITUDE QUESTIONNAIRE (BASIC AND AIT)

INSTRUCTIONS: The purpose of this questionnaire is to obtain information from you regarding training, working and living while in the Army's Basic Training and Advanced Individual Training (AIT) program. Your answers will help the Army to determine what conditions are in need of improvement, and will assist the Army in determining the actions they must take to improve training and the quality of life for new soldiers in the Army. Your honest opinions are, therefore, essential.

We have no need to know who you are personally. No effort will be made to identify either you or your unit. DO NOT WRITE YOUR NAME, SOCIAL SECURITY NUMBER, OR UNIT on either the questionnaire or the answer sheet.

Each question should be answered by circling the letter on your answer sheet which is next to the answer which best describes your feelings. See sample question below:

SAMPLE QUESTION: 3. How old are you?

a. 17
b. 18
c. 19
d. 20
e. 21 or older

If you are 19 years old, you should circle the letter c on your answer sheet for question 3, as has been done below, since the letter c corresponds to your correct age of 19 on the questionnaire.

QUESTION NUMBER	RESPONSES (CIRCLE ONE)				
01	a	b	c	d	e
02	a	b	c	d	e
03	a	b	c	d	e
04	a	b	c	d	e

If you have any questions, please ask the questionnaire administrator for assistance. You will have 30 minutes to complete the questionnaire. You will all turn in your answer sheets, and leave at the same time. Do not turn the page and start to work until instructed to do so.

C. Anonymity for Respondents

1. Factors to be Considered

There are several factors to be considered when deciding whether to require the respondent's name or other identifying information on a questionnaire. Some of the factors are supported by research, while others are not.

a. If the respondents supplied their names, they are aware that they can be identified and called back. If respondents do not have to give their names or similar information, most will believe that they cannot be identified and called back for any type of accounting after their questionnaires have been collected.

b. The perception of anonymity seems to depend not only upon whether respondents give their names, but also on the conditions under which the questionnaires are administered. For example, paper-and-pencil questionnaires are more anonymous than structured interviews.

c. The effects of anonymity seem to be related to the content of the questionnaire. This is particularly true when information on sensitive areas is collected. For general attitudes, it may not matter.

d. The effects of anonymity may also depend upon who administers the questionnaire, and the circumstances under which it is administered. Responses may be distorted when respondents are identified and under high threat.

e. Respondents may be more lenient when rating other personnel if they think they will be identified.

2. Implications of the Privacy Act of 1974

If the experimenter, test officer, or questionnaire writer desires to obtain certain types of personal information from a respondent, the federal Privacy Act of 1974, in turn, requires that certain information first be given to the candidate respondent. One may use DA Form 4368-R, 1 May 75 for the purpose of communicating this information to the respondent. The form is shown filled out on page X-C 4. In this particular example, the research questions dealt with attitudes toward respondents' treatment in the Army.

A second example, Figure X-C-1, illustrates a more compact format. The same elements of information called for by DA Form 4368-R have been communicated; it's just that that form was not used.

A privacy act statement is not necessarily required as a part of all questionnaires that are administered to Army personnel. It is not necessary where only the personal information listed below is being requested. For example, no invasion of privacy is involved where soldiers are asked to evaluate some new/revised weapon, equipment, or organization regarding effectiveness and/or acceptability, and to answer any of the 12 items listed below. The collection or release of the following information does not require the consent of the respondent:

a. Grade.

b. Date of birth.

c. Date of rank.

d. Salary.

e. Present duty assignment.

f. Past duty assignments.

g. Future assignments (approved).

h. Unit and/or office address.

i. Unit and/or office phone number.

j. Source of commission.

k. Military and/or civilian education.

l. Promotion sequence.

Data collection procedures that guarantee anonymity are desirable for surveys. If the research methods cannot guarantee anonymity, then confidentiality of the data is to be protected. For operational test and evaluation research, participants should be informed that the data cannot be kept confidential. Surveys requiring the names of participants can have records coded as soon as possible. The key to the code can be stored for limited access to protect confidentiality. (See American Psychological Association (1982). Ethical principles in the conduct of research with human participants. Washington, DC.)

If you have any questions concerning the Privacy Act of 1974, you may obtain additional information from the ARI Field Unit. All questionnaire respondents must be advised of the requirements of the Privacy Act of 1974 when any of the 18 types of information listed below are being requested. This information can only be obtained from an individual on a voluntary basis. The release of any of the information listed below requires the prior and informed consent of the individual.

a. Name.

b. Social Security number.

c. Home address.

d. Home phone number.

e. Home of record.

f. Financial transactions.

g. Character quality.

h. Efficiency ratings.

i. Conduct ratings.

j. Legal affairs.

k. Religious preferences.

l. Number of allotments.

m. Amount of allotments.

n. Medical history.

o. Criminal history.

p. Fingerprints.

q. Voiceprints.

r. Photographs.

DATA REQUIRED BY THE PRIVACY ACT OF 1974
(5 U.S.C. 552a)

TITLE OF FORM	PRESCRIBING DIRECTIVE
	AR 70-1

1. AUTHORITY

10 USC Sec 4503

2. PRINCIPAL PURPOSE(S)

The data collected with the attached form are to be used to research purposes only.

3. ROUTINE USES

This is an experimental personnel data collection form developed by the U.S. Army Research Institute for the Behavioral and Social Sciences pursuant to its research mission as prescribed in AR 70-1. When identifier (name or Social Security Number) are requested they are to be used for administrative and statistical control purposes only. Full confidentiality of the responses will be maintained in the processing of these data.

4. MANDATORY OR VOLUNTARY DISCLOSURE AND EFFECT ON INDIVIDUAL NOT PROVIDING INFORMATION

Your participation in this research is strictly voluntary. Individuals are encouraged to provide complete and accurate information in the interests of the research, but there will be no effect on individuals for not providing all or any part of the information. This notice may be detached from the rest of the form and retained by the individual if so desired.

FORM Privacy Act Statement · 26 Sep 75

DA Form 4368-R, 1 May 75

Figure X-C-1

An Example of a Privacy Act Statement

11B/C GRADUATE FIELD SURVEY
(Prescribing Directive: AR 600-46; TRADOC Ltr dtd 29 Aug 75)

INFORMATION PRIVACY ACT STATEMENT

1. Authority: 5 USC 301, 10 USC 3012, Authority for the Secretary of the Army to Issue AR's; 44 USC 3101, Authority for Collecting Necessary Data.

2. Principal Purpose: To collect data to evaluate the effectiveness of individual training received prior to joining one's initial unit of assignment.

3. Routine Uses: The data collected with this form are to be used for research purposes only. They will not become a part of any individual's record and will not be used in whole or in part in making any determination about an individual.

   The identifiers (name or Social Security Number) are to be used for administrative and statistical control purposes only. Full confidentiality of responses will be maintained in the processing of these data.

4. Mandatory or Voluntary Disclosure and Effect on Individual Not Providing Information: Voluntary - Your participation in this research is strictly voluntary. Individuals are encouraged to provide complete and accurate information in the interests of the research, but there will be no effect on individuals not providing all or any part of the infomation.

   This notice may be detached from the rest of this form and retained by the individual answering the questionnaire if so desired.

D. Motivational Factors

This section considers the effects of lack of motivation, and some ways of providing a desirable level of motivation to respondents during the questionnaire administration process.

1. Effects of Lack of Motivation

Generally, the results of any study will suffer distortion if those to whom the questionnaire is distributed are not sufficiently motivated. If they have the choice, they will not respond at all. If they do have to respond or are just minimally motivated, they may omit items, make patterned or random responses, or just generally respond poorly. As a result, the reliability and validity of the responses will be decreased, and the results of the study would lead their reader/user into some degree of error.

2. Ego Involving Potential Respondents in the Study

There are a number of ways that motivation can be increased by ego involving potential respondents. Some of the ways are given below:

a. The special role of the respondent in the study can be emphasized.

b. Responsibility can be stressed when it is appropriate to do so.

c. The wording of cover letters, if used, affects ego involvement. Help may sometimes be requested on the basis of appealing to the self-interests of the respondent. There is evidence that this type of appeal helps most with less educated respondents.

3. Stimulating the Return of Remotely Administered Questionnaires

Obviously, whatever involves the egos of potential respondents in a study also stimulates the return of remotely administered questionnaires, such as those distributed by mail. Other ways of stimulating the return or response rate are:

a. Return rates may often be significantly improved when a letter is sent in advance notifying the potential respondents that they will receive a questionnaire and their help is needed in filling it out.

b. Stamped and return addressed envelopes can be sent with the questionnaire. There is evidence that this does increase response rate.

c. There is contradictory evidence about whether short questionnaires are returned more frequently than longer ones, but one would probably believe this to be true.

d. Follow-up reminders can be sent to those who do not promptly return their questionnaires. There is some question, however, regarding how much such follow-ups increase response rate. At times, it may not be cost effective, so maybe the decision should be a function of whether or not the initial return rate was adequate.

e. Telephone interviews and face-to-face interviews generally have a higher response rate than mail surveys.

f. Response rate for telephone interviews can be increased by changing the format from what would be used in a face-to-face interview. Select fewer items and items which are shorter in length for telephone interviews to reduce telephone disconnects.

g. Nonresponse for items following a branch may increase the overall item nonresponse rate, especially for mail surveys. If branching can be avoided, this may increase item response rate.

4. Use of Incentives

The evidence has been mixed regarding the extent to which motivation is increased through the use of incentives. Incentives may include money, time off, special privileges, etc. Generally, however, it is agreed that incentives usually help increase the response rate with remotely administered questionnaires.

5. Other Motivational Factors Related to Questionnaire Administration

Many additional motivational factors related to questionnaire administration can be noted or inferred from other sections in this manual. Some of them are:

a. Respondents often have preferences for certain item formats, although sometimes such preferences may not offer any advantage in terms of reliability and validity. Some subjects prefer rating scales to forced choice items. With forced choice, some like the option of indicating the degree of applicability of each statement. Some do not like forced choice Q-sort (see Section IV-H). Some prefer multiple category to two category options. In some studies, Likert scales have been preferred to behavioral scales. Behaviorally Anchored Rating Scales have been preferred to Mixed Standard Scales, etc. These preferences may relate to familiarity of the respondent with given item types. There is not much that the questionnaire designer can do about such preferences, except to note that they exist.

b. Researchers in recent years have explored the cognitive complexity of respondents to match them to formats which are cognitively compatible. There have been problems with replication for this research.

c. Motivation may be increased by offering feedback of study results to the respondent.

d. Every effort should be made to praise the respondents or potential respondents, to the extent that it is reasonable.

e. Long, vague, or boring questionnaire sessions should be avoided, since it will decrease respondent motivation to continue attending and providing "best" responses.

f. Questionnaire administration sessions should not be scheduled when there are conflicts with other activities of greater interest to the respondents. Nor, in general, should they be scheduled very early or very late in the day.

g. Volunteers are usually more motivated to fill out questionnaires than are nonvolunteers. However, their replies may be more biased.

h. When respondents are told that they may leave as soon as they have completed the questionnaire, they usually do a much more hasty and unsatisfactory job than when they are given a specific time for completion, and are told that they cannot leave until the time period is up.

i. See Chapter XIV about the behavior of interviewers.

E. Administration Time

Little is known about the effects of questionnaire administration time on respondents' motivation, or of the effects of setting time limits for completing questionnaires. The questionnaire administration period should generally have been determined in advance by pretesting. There will be some variability in the length of time taken to complete a questionnaire. There is remarkable consistency among those who are sincere in attempting to do an accurate and complete job of answering all questions.

When a questionnaire is administered to a group of respondents, the instruction should emphasize that all respondents will be given plenty of time to answer the questions. As indicated earlier in X-D 5 h, the instructions should not tell the respondents that they can leave as soon as they have finished the questionnaire. Many will then cut short their efforts to answer the questions. There is little hope of obtaining carefully considered evaluative responses on a questionnaire if the respondents knows that the faster they finish the questionnaire, the sooner they will be able to go home.

Questionnaire administration time is obviously related to questionnaire length, which is the topic of Section IX-C.

One should try to determine empirically the maximum time needed to complete a given questionnaire. If the questionnaire is group-administered, the maximum time for the slowest respondents should usually be used in scheduling the administration of the questionnaire.

F. Characteristics of Administration

Little has been established in the research literature about how the characteristics of questionnaire administrators affect the overall process with nonremotely administered questionnaires. The following items may be noted:

1. In most cases, it is felt that the sex of the administrator has no effect on the responses received. There may, however, be certain motivational effects.

2. The military rank of the administrator may have an effect on the respondent, but no research has been performed to examine this.

3. Any effect that the race of the administrator has on the respondent may also be a function of the content material of the questionnaire, e.g., race would be expected to influence responses on a race relations questionnaire more than on a questionnaire dealing with rifle comparisons. The effects should probably be viewed as the result of interaction between administrator and respondent characteristics, and the questions being asked.

4. Implications exist for biasing survey results whenever surveys incorporate face-to-face interviewing with individuals from different ethnic backgrounds. Items with racial content used in a questionnaire are especially sensitive to such biasing.

5. See Chapter XIV about the influence of an interviewer on the interviewee.

G. Administration Conditions

Questionnaire administration conditions obviously cannot be controlled with remotely administered questionnaires. With group-administered questionnaires, the following guidance is offered:

1. Administration conditions should be provided which are most appropriate to the particular type of respondent completing the questionnaire.

2. Administration conditions have an effect on questionnaire responses. For example, different responses may be obtained if the questionnaire is filled out in a group situation on the job rather than individually at home.

3. When personnel are being rated, different ratings may be obtained, depending on how acquainted the rater and ratee are.

4. For Army field test evaluations, the circumstances under which questionnaires must/can be administered will vary rather widely. There may be times when no writing surface(s) or pencils are available; clipboards and pencils should be supplied if this problem can be anticipated. If the needed materials cannot be brought to the respondents, then arrange to move them to a place where the materials and other environmental conditions are satisfactory.

5. Respondents should be required to give their answers without being influenced by other respondents. Achieving this requires respondents to be somewhat separated and/or to have the administrator(s) watching them. Simply instructing them not to consult with each other is usually not sufficient.

H. Training of Field Test Evaluators

An extended discussion of the training of raters and other test evaluators is not undertaken in of this manual. The following suggestions, however, can be offered about the general training of the Army field test evaluators. See Section X-B regarding questionnaire administration instructions.

1. Impress on test evaluators that they are supposed to answer the questionnaire based upon what they observe in the test. Stress the need for evaluations based only upon what was seen during the test exercise, regardless of any personal feelings or knowledge of concepts or equipment as might exist in a true combat environment (except in special instances where this is specifically asked for). To help identify and reduce prejudgment, a broad question might be included to permit the evaluators to express any biases they may have. It may be a question such as "Based on your personal experience, do you feel the "DPST" is a useful approach to real daily problems, i.e., outside a test exercise environment?" Such a question would permit the evaluators an outlet for preconceived opinions and attitudes which otherwise would color their view of the events observed during the exercise. On the other hand, in some situations the evaluators might feel it necessary to defend their personal judgment by biasing their answers to the remaining question answers!

2. Stress the importance of evaluators to the success of the test. Perhaps briefly indicate some actions which have been taken to implement concepts supported by evaluative data from previous tests.

3. Permit evaluators (particularly after the pilot test) to sound off about the forms and their perceived inadequacies, regardless of how unreasonable these complaints might be. The goal is to have all evaluators answering questionnaires understand that they are active and important contributors rather than just a means of satisfying some obscure test requirement.

4. Examine all turned-in questionnaires to ensure that they have been filled out and understood. This procedure should continue throughout the entire series of tests.

5. Stress the notion that complete honesty and objectivity is needed. Sometimes evaluators try to please the test sponsors, to the detriment of the test.

6. Indicate to evaluators, perhaps on the top of all questionnaires or verbally, that they may make marginal note clarifications concerning their scale value selection for any rating question. This will increase posttest accuracy in determining questions which are scaled awkwardly or unclearly stated. This is particularly crucial during the pretesting or pilot test. Notes should be made regarding question structure immediately as they occur to the evaluator or the difficulty is likely to be forgotten.

7. Prior to having the evaluators complete questionnaires, ask all or a sample of randomly selected evaluators to orally describe to the other evaluators what they believe each question is asking. This procedure will reduce differences between judges because of varying semantic interpretations. By the time of the actual exercise, all evaluators should generally agree, for example, on the meaning of "command and control effectiveness," "fire power potential," etc. If this is done, the criteria will have mutual acceptance. This procedure is also useful during the pretest to assist in the selection of item wording that will be understood by the respondents.

8. Evaluators should be forewarned about biases such as the halo effect, central tendency, and others discussed in Chapter XII. If it is explained to the evaluator that these are common biases to which we are all subject, the evaluators will be better able to consider the fairness and accuracy of their observations. Training to reduce rating errors is especially effective when the training is extensive, and allows evaluators to practice. Evaluator experience with the questionnaire may improve rating accuracy. Short training programs may have little impact on rating quality. To train evaluators, effective training should include observational techniques in conjunction with written performance observations between rating periods.

9. The independent, non-collaborative, evaluation of each question should be stressed.

I. Other Factors Related to Questionnaire Administration

Some other factors related to questionnaire administration that have not been discussed in other sections of this manual are addressed below:

1. Respondents may at times be influenced by the title of the questionnaire. The word "test" should not be used in a title of a questionnaire as it may imply that it is a test of the respondent's knowledge.

2. A problem with Army field test evaluations concerns undue influence by the questionnaire administrator. It is sometimes necessary to use line officers from the units of the test subjects as questionnaire administrators. When outside administrators are used, they must be carefully instructed to make no comments whatsoever regarding their personal opinions of the items being evaluated. An offhand comment by a company commander administrator to his/her company regarding the "goodness" or "badness" of a piece of equipment or concept being evaluated can exert an influence sufficient to distort the results significantly from what they would otherwise have been.

3. The manner in which test subjects are selected and utilized in operational tests may affect the manner in which they respond to questionnaire items. For example, separate groups with no prior experience with either the test system or the current standard system could evaluate each system. This would exclude pretest biases, but test subjects would have no basis to compare the two systems. Alternatively, the same group of test subjects could use both systems in rotation. However, this procedure may result in a bias for or against one or both systems as a function of which was used first. In this respect too, personnel having extensive prior experience with a current standard system may introduce their pretest biases for or against that system when it is being evaluated against a candidate replacement system. The consequence of such considerations is that the type of system evaluation intended will govern the way evaluators and/or test subjects are selected and utilized. The methods of selection and utilization will influence the way questionnaires must be designed, and in turn suggest the types of problems likely to arise.

Chapter XI: Pretesting of Questionnaires

A. Overview

Even the most careful screening of a questionnaire by its developer or by questionnaire construction experts will usually not reveal all of its faults. Pretesting is an important and essential procedure to follow before administering any questionnaire. Its purpose is, of course, to find those overlooked problems and faults that would otherwise reduce the validity of the information obtained from the questionnaire responses. However, just any pretest will not do. One must know how to pretest the items and what to look for.

Some guidelines for pretesting questionnaires are given in this chapter. Pretesting may seem to some uninformed individuals to be a waste of time, especially when the author may have asked several people in his/her own office to critique the questions, or perhaps even asked a questionnaire specialist to critique it. However, pretesting is an investment that is well worthwhile. It is crucial if the decision that will result from the questionnaire is of any importance.

B. Guidelines for Pretesting Questionnaires

1. Before a pretest is conducted and a questionnaire is constructed, hypotheses and questionnaire items are developed. The hypotheses are presented to a group of individuals who are subject matter experts. The group performs a preliminary assessment of the hypotheses and items. Modification may be required regarding the hypotheses and/or questionnaire items.

2. Initially, open-ended questions are established and placed into a logical sequence. Pretesting may provide information that can be used to convert open-ended questions to multiple choice questions to facilitate data reduction and analysis. Instructions are developed to accompany the questionnaire, and they are included as part of the pretest. If branching is used, it should be kept to a minimum.

3. It is important that the respondents employed in pretesting be representative of the eventual target respondents. For example, if infantry enlisted men will perform in a test and then take the questionnaire, it should not be pretested with respondents who are armored officers; even infantry officers would not be satisfactory.

4. The pretest is more useful if it is conducted by someone who knows the operations to be performed in the test and who also knows the subject matter that the questionnaire covers. It is best if the question writer is knowledgeable about these operations and conducts the pretest.

5. Early versions of questionnaires may contain instructions, item stems, response alternatives, and item ordering that are confusing to respondents. It is possible that more than one pretest will need to be conducted. Some researchers have been known to conduct up to six or more pretests.

6. Interview and pretest some of the pretest respondents one at a time or in a group. Ask each respondent to read each question and explain its meaning. Also ask them to explain the meaning of the response alternatives, and to make their choice. Ask the respondent to explain why a particular choice was made. The respondents' answers will frequently reveal incorrect assumptions and possible rationales that the question writer never dreamed possible. They will also help to identify lack of understanding of particular words, vague or ambiguous phrases, ill defined or loaded questions, inadequate space for recording answers, inappropriate sequencing of items, etc.

7. One good technique for pretesting is to have respondents complete the questionnaire. A discussion can then be had where respondents read each question aloud and then tell you what it means. Any difficulties at all should be a cause for concern and revision. Pretest methodology can be strengthened if discussions are tape recorded, and suggestions for modifications are systematically coded. This is especially useful when pretesting a questionnaire that will be administered by interview.

8. During pretesting, the respondents should be encouraged to make marginal notes on the questionnaire regarding sentence structure, unclear questions or statements, etc. Pretests will provide a good idea as to the length of time it takes to complete the questionnaire.

9. When attitude questions, especially, are being pretested, individuals who may hold minority views should be included. This will help identify loaded questions.

10. Pretests for the selection of verbal anchors are valuable in building rating scale content validity and reliability. Rather than employing anchors which seem appropriate, the anchors used in the final scales should be selected as a result of analyses of pretests of respondents similar to those who will be participating in the final test.

11. While pretesting a questionnaire, a high proportion of respondents giving no response or a "Don't know" response should be a cause for concern. However, a low number of "Don't know" responses (especially for multiple choice items) does not guarantee that the question is good.

12. After pretesting, each question should be reviewed and its inclusion in the questionnaire justified. Questions that do not add significant information or that largely duplicate other questions can profitably be eliminated. Quantitative item reduction techniques will depend on the type of scale that is being used, e.g., Thurstone, Likert, Guttman, etc. A discussion of quantitative item reduction techniques is outside the current scope of this manual. Army personnel may check with the Army Research Institute Field Unit closest to them for help in this area.

Chapter XII: Characteristics of Respondents That Influence Questionnaire Results

A. Overview

This chapter discusses some characteristics of respondents that influence questionnaire results. It therefore identifies some of the principal sources of error in the reporting of observations and/or the evaluation of performance in, for example, operational Army field tests. Additional research is required, however, to determine their relative contributions to error variance.

Sections XII-B and C present a discussion of various biases, response sets, or other sources of error. There is some confusion in the literature regarding the use of these terms, but they are similar. A bias is: a tendency to deviate from a true value; a tendency to favor a certain position or conclusion; or an attitude either for or against a certain unproved hypothesis which prevents an individual from evaluating the evidence correctly. A response set or response bias refers to the tendency of a respondent to answer questions in a particular way almost independent of the content of the questions. An error is simply a mistake or departure from correctness.

Section XII-D addresses the effects of attitudes of respondents on questionnaire results, while Section XII-E considers the effects of demographic characteristics on responses.

One of the main purposes of this chapter is to alert the questionnaire designer to some of the characteristics of respondents that influence questionnaire results. There are ways that some of the biases and errors can be controlled, but not all of them. And there appears to be no easy way of detecting the influence of a response set nor of neutralizing it. More detailed identification and control methods are areas of needed further research.

B. Social Desirability and Acquiescence Response Sets

Social desirability is a response set where persons answer according to the norms they believe society supports. It is the tendency to agree with items the respondents believe reflect socially desirable attitudes in order to show themselves in a better light. Acquiescence response set is the tendency to consistently agree, to say "Yes," or to say "True." It is a general tendency to assent rather than dissent. Although there have been a number of studies about each, a detailed discussion of them is beyond the scope of this manual. (See P-77-2, Questionnaire Contruction Manual Annex, Literature Survey and Bibliography; and P-85-j, Questionnaires: Literature Survey and Bibliography.) Some comments about each are presented below.

1. Social Desirability Response Set

a. Social desirability response set seems to operate whenever the respondent has the opportunity to respond in terms of it. Some believe that its effect is so powerful that respondents would not tend to deviate from social norms in their answers even though their behavior denied what they said.

b. Several authors have identified respondents with a high social desirability response rate. They found these respondents to give more true responses to neutral items, to be more susceptible to social pressures, to more likely be introverts, and to score higher on a "lie" scale.

c. Faking or responding with socially desirable answers which are not true is part of the response set.

d. Anonymity fails to eliminate the social desirability response set.

e. The forced choice instrument format has been studied for its susceptibility to social desirability response set, a factor it was intended to control. Some authors found the forced choice method minimized the effects of social desirability, while others think the factor still needs additional control. One study concludes that in comparing different forced choice formats, ambiguous items tend to be freer of social desirability response set than positively or negatively worded items. In any case, the evidence indicates that the social desirability problem is sometimes less in forced choice formats than in other item types such as graphic rating scales. Forced choice formats may or may not reduce bias.

f. Card sorts also need control to eliminate social desirability bias.

g. Respondents may be confounding trait dimensions with response alternatives on clinical instruments. There is some evidence that respondents have a stronger tendency to select response alternatives opposite in desirability when a socially undesirable response alternative is presented first.

h. Procedures have been developed for controlling or balancing social desirability by using loaded items in the questionnaire and then adjusting the respondent's score. The social desirability score from the loaded items can also be correlated with each of the other items on the questionnaire. The responses on those items with a statistically significant correlation can then be corrected by moving the response one or more steps from the socially desirable response to give a more accurate result.

2. Acquiescence Response Set

a. The acquiescence response set is defined as a behavioral attitude by the respondents to agree and accept, even if they must alter their original opinions to do so.

b. The acquiescence response set seems to operate especially when statements are in the form of plausible generalities.

c. The response set may occur more with difficult than with easy questionnaire material.

d. Acquiescence response set may be a personality trait.

e. There is a concern that social desriability and acquiescence response sets may be related in such a way that an individual with a tendency toward conformity will consistently reflect both biases.

f. Controls for acquiescence response set have been researched. Stating the question stem in a neutral manner may help minimize acquiescence. The effects of acquiescence response set may also be partially controlled by using two alternate questionnaire forms with the question stated positively on half of the forms and stated negatively on the other half. The balancing of scales (e.g., equal number of positive and negative points) may also be of value in counteracting acquiescence.

C. Other Response Sets or Errors

This section notes a number of other response sets or errors of which the questionnaire developer should be aware.

1. Error of Central Tendency

Some respondents tend to avoid endpoints on a scale, and pick a middle value regardless of their true feelings. It may be more common when the respondents are not very familiar with whatever they are being asked to rate. It may be counteracted by adjusting the strength of the response alternatives so that there are greater differences in meaning between alternatives near the ends of the scale than between alternatives near the center.

a. In one study, responses tended to be toward the center of the scale when item length increased (more than 17 words). Respondents selected response alternatives toward the positive end of the scale when item length was short (less than 17 words). Items may be ambiguous to the respondent when they are long and negatively worded. This appeared to influence respondents to rate items toward the mid-range of the scale.

2. Extreme Response Set

On the other hand, some individuals tend to consistently select exaggerated choices for positions. It can be recognized when a respondent makes a pattern of answers which tend to be unevenly distributed toward one or both ends of a scale. Research indicates that this response set may be a personality characteristic.

a. Research evidence indicates that positively worded items receive higher mean responses than negatively worded items. There is the possibility that respondents prefer or agree with positively worded items, and rate them higher.

b. For cross-cultural survey research, there is some evidence that response style may vary from country to country. One study concluded that there was a tendency by respondents in the Philippines to use a positive response style, and by respondents in Italy to use a negative response style.

3. Halo Effect

Halo effect was originally defined as a tendency when one is estimating or rating a person with respect to a given trait, to be influenced by some other trait or by one's general impression of the person. It is, however, also applicable to ratings of other than people. For example, if field test evaluators know that a particular weapon system did well in one phase of a test, they may be influenced to give high ratings to the system in later test phases - even when the system performs poorly.

a. Most studies of ways to control halo effect have dealt with ratings of traits of personnel by other personnel, a matter not of great concern in this manual. The forced choice technique and Mixed Standard Scales minimize halo effect in some situations. Ratings will also be less distorted if questionnaire items are constructed so as to relate to clearly observable aspects of behavior which do not overlap. It is doubtful that the influence of halo effects can be completely eliminated from the responses to any questionnaire.

b. Behavioral scales such as Behaviorally Anchored Rating Scales (BARS), Behavioral Expectation Scales (BES), and Mixed Standard Scales (MSS) have been developed to measure performance. There is evidence that the use of behavioral scales in conjunction with intensive training can reduce halo error. This combination of behavioral scale and training appears to be more effective than graphic rating scales, trait scales, and Likert scales in reducing halo error. The length of the training session appears to influence whether halo error will be reduced. Training sessions of 5-minute duration have had little impact on the quality of ratings. Training sessions of 3-hour duration were found to reduce halo error. Intensive training sessions may not reduce other types of rating errors even though they tend to reduce halo error.

4. Leniency Error

Leniency error refers to a general, constant tendency for a rater to rate either too high or too low in the direction of being too generous. It appears similar to halo effect, except that it is independent of the trait or factor being rated. Some raters have an opposite tendency to rate too severely. In one study, respondents rated Likert scales with less leniency error than they rated behavioral scales. These findings may not be consistent across studies. In large groups of raters, the opposite tendencies should balance out.

5. Logical Error

Logical error is also similar to halo effect. It is due to the fact that raters are likely to give similar ratings to traits or items that seem logically related. For example, field test evaluators may know that a counterattack was extremely successful; they may therefore reason that command and control was also very effective and should receive an equivalent high evaluation because a successful counterattack is a function of good command and control. Such reasoning assumes a dependence which may or may not be true. Logical error may be avoided in part by asking for judgments of objectively observable actions or behavior.

6. Proximity Error

Proximity error occurs when, due to the ordering of questionnaire items, the answer to one item results in an answer to a subsequent question being substantially changed from what it would otherwise have been. Little is known about its influence in field test situations; most research in this area has concerned the rating of personality trait variables.

7. Contrast Error

Contrast error refers to a tendency of raters to rate others in the opposite direction from themselves in regard to a trait. Little research has been done on this source of error.

8. Feedback Bias

Research shows that if observers are informed of experimental hypotheses, and if they receive daily feedback indicating how well their data support the hypotheses, they will tend to report data supporting those hypotheses - even when the reverse is true! This bias does not seem to occur, however, when observers are informed only of the experimental hypotheses with no follow-up. Taking precautions to assure high levels of observer accuracy minimizes the bias.

D. Effects of General Attitudes of Respondents

Limited research has been conducted upon how the attitudes of a respondent influence questionnaire results. The following, however, should be noted:

1. Respondents at times base their ratings not on what is observed but on what they believed prior to the observation. Beliefs and opinions may affect results.

2. It is generally believed that judges used as part of the process of determining scale values can rate items without being influenced by their own attitudes. There is also some evidence to the contrary.

3. Unstable or changing responses to questionnaires may be caused by shifts in the mood of the respondent, relative values among the possible choices, and the degree of interest present in the question.

4. As questions become more ambiguous, responses normally become more influenced by attitudes.

5. It may be desirable to revise a questionnaire when norms of groups differ greatly from those with whom the questionnaire was pretested or previously administered.

E. Effects of Demographic Characteristics on Responses

Demographic characteristics have been shown to influence questionnaire results. Similarities of such variables among respondents often tend to be related to a response pattern. These variables include: age, religion, sex, intelligence, marital status, parenthood, socioeconomic class, nationality, race, urban or rural residence, income, rank and experience. Questionnaires should, therefore, be designed with the respondents background in mind. When there is a suspicion that demographic characteristics may affect response, the data should be analyzed by type of respondent.

1. Research indicates that the racial background of survey interviewers does not seem to affect survey results when the questions do not deal with racial stereotypes and are not threatening. For most questionnaires administered by interview, it would be possible to assign interviewers of different racial backgrounds regardless of respondents' racial backgrounds.

2. Racial background has been known to influence rating errors on performance measures. However, this phenomenon has not been observed consistently.

3. Survey items which tend to be most sensitive to differences in response pattern by gender are those dealing with sex role stereotypes. Items that are relevant to technical background experience which females may not have, may yield gender-related response patterns.

4. It was hypothesized in one study that females would rate items according to social desirability. It was suggested that females have a greater need for social approval, that they are more impulsive than males, and that this would be reflected by male/female differences in rating. The results indicated that there were no significant differences between ratings by gender. Some studies have found differences in rating by females, while other studies have not. The overall effect of gender differences in response pattern has little support. Rating characteristics identified by gender are usually not enough to explain rating differences. Other variables must be taken into account as well, such as education, race, age, etc.

5. Respondents with low levels of education may be confused by items constructed in absolute terms. This may result in rating the items with an inconsistent response pattern. One should review questionnaire items to ensure that the content is not ambiguous to respondents with a low level of education.

6. Survey items which request an opinion regarding an obscure topic area may elicit a "Don't know" response by respondents with higher levels of education. Respondents with higher levels of education seem to be more willing to admit they do not have knowledge of obscure topics. For this type of question, respondents with less education have a tendency to give an opinion. Respondents with low levels of education do not appear to admit they don't know, but instead select a response alternative to represent their opinion.

7. For questions which are not obscure, the "Don't know" response alternative may be selected most frequently by respondents who have the least amount of education. Individuals with less education appear to be the most influenced when a "Don't know" response alternative is included.

8. The age of respondents does not appear to influence their ability to use different types of rating scales. However, the educational level of respondents may affect the way in which different scales are rated.

9. The content of some items may be related to the historical perspective of different age groups. For such items, the responses may be associated with different response patterns according to the age of respondents.

10. Nonresponse to an entire survey or to specific items in a survey remains a threat to the validity of survey results. Research indicates that nonresponse rates are sometimes associated with age of respondents. Item nonresponse rate may be reduced by eliminating branching from surveys that include respondents approximately 60 years and older.

11. When surveys are conducted by interview, it is important to be sure that older interviewers (about 50 or over) are following the standard format and interview guide. Of course, all interviewers need to conduct standardized interviews. One study indicated that older interviewers made more errors by conducting the interview in a nonstandardized way. Possibly, the interviewers felt that their years of experience afforded them the opportunity to probe questions more thoroughly, and to somewhat modify the interview guide as they progressed through the interview.

Chapter XIII: Evaluating Questionnaire Results

A. Overview

An extended discussion on evaluating questionnaire results is currently outside the scope of this manual on questionnaire development. However, Army personnel may check with the Army Research Institute-Field Unit closest to them for help in the areas of coding and data analyses. There are some factors relating to the evaluation of questionnaire results that should be noted since they may influence how questionnaires are designed and developed. Section XIII-B considers the scoring and coding of questionnaire responses, and Section XIII-C contains some notes about data analyses.

B. Scoring Questionnaire Responses

1. Practical Considerations

a. Both time and money can be saved by planning the questionnaire in line with scoring and tabulation requirements. The phrasing of questions and their sequencing and layout affect tabulation time. For example, it is advantageous to have data coded and entered for analysis directly from edited questionnaires. Questionnaires consisting of only closed-end items will have a lower level of error for data entry than open-ended items. This is a more cost-effective approach. However, there are some drawbacks such as greater difficulty in verifying the coding and greater data entry time than when using a coding sheet.

b. A decision should be made ahead of time regarding whether the data will be tabulated by hand or machine.

c. Response alternatives should be precoded whenever possible. Codes for open-ended items are more difficult to construct than codes for closed-end items. To develop open-ended item codes, list out possible responses to the item. Pretest the questionnaire to classify responses to open-ended items. Construct a classification system and code. Pretest the code and revise as necessary. Develop a separate code for responses that were not possible to fit into the classification system above.

d. Codes need to be developed which guide coders in assigning code numbers to each answer. This includes the following: codes for missing data for item nonresponse, codes for item responses that are uncodable due to poor respondent performance, and a code for the "Don't know" response alternative.

e. Code books are constructed to define, clarify, and amend codes used during the coding process. Codes that have caused difficulty for the coders should be noted, such as classification systems and codes for open-ended items. Coders require training on specifics of the classification system and codes used for the study, and for the general principles of coding.

f. Since it does not seem to matter if items are scrambled or in blocks according to content, blocking may be preferred due to greater hand scoring ease.

g. Telephone surveys now use Computer Assisted Telephone Interviewing (CATI). These systems are still in experimental stages, and they require extensive programming. Items are read off the CRT screen, and telephone interviewers type respondent answers into a terminal for direct data entry.

h. See Section IX-E regarding the use of answer sheets.

2. Other Considerations

a. There may be a justification for scoring rating scale items dichotomously according to the direction of response. It is sometimes done when bipolar scales are analyzed in terms of the proportion of responses in either direction of the basic dichotomy. The justification is based upon results that seem to indicate that composite scores reflect primarily the direction of responses and only to a minor extent their intensities.

b. One investigator found that many Likert-type rating scales consisting of 2 through 19 steps may be collapsed into two or three measurement categories for analysis with no lack of precision.

c. When working with paired comparison items with a "No preference" option, the "No preference" responses can often be either divided proportionate to the preference responses, or disregarded altogether. The basis for this suggestion is that respondents who claim neutrality appear to exhibit the same preference patterns as those who express a preference.

d. By using any one of several methods of scoring or transforming self-rating scale raw scores, it is usually possible to approximate dichotomous forced choice results with considerable saving in administration time, and a small gain in test-retest reliability.

e. Investigators sometimes use intensity scores as well as rating scale content scores. One way of obtaining an intensity score is to follow each question with the query, "How strongly do you feel about this?" A second way involves weighting extreme responses (positive and negative) as 2, moderate responses as 1, and neutral responses as 0. These weights can then be summed for an intensity score.

C. Data Analyses

A detailed discussion of data analyses is beyond the scope of this manual; however, some basic data analysis issues have been mentioned in related chapters. Additionally, the following points are also noted:

1. Analyses of questionnaire responses are chiefly of two types: summary tabulations and statistical analyses. Tabulations are used primarily for the presentation of results. Statistical tests are used to determine whether the differences in the results are significant. Statistical literature is available which presents numerous tests usable in such analyses.

2. As part of the questionnaire development process, tentative (dummy) analysis tables should be developed to assure that the data to be obtained are appropriate.

3. Weights can be assigned to questionnaires when there is a probability that the selection of respondents is not representative of the population as a whole. For example, a sample distribution drawn from a list of service personnel receiving training, and enrolled in various courses, may result in unequal probability sampling. Since the subjects may be enrolled in more than one course, the more courses they take, the greater the chance they will be selected into the sample.

   Weights are also used in making adjustments for total nonresponse and in poststratification. They are able to assign greater importance to some sampled elements than to others in the data analysis. Poststratification conforms the sample distribution to the known population distribution. The sample distribution is adjusted across the strata. This is useful when the population is known, but the stratified sample elements cannot be determined at the selection stage. In such situations, prior stratification is not employable, although poststratification may be applied later. When a sample is weighted to a known population, it will adjust for the sampling fluctuations, as well as for nonresponse. For example, if nonresponse is higher for a specific age group, the sample will conform to the known age distribution when weighted. The development of weights is a difficult task. Standard computer programs for weighted data can be applied in data analysis.

4. Four kinds of measurement scales have been identified: nominal, ordinal, interval, and ratio. Appropriate statistical analyses are associated with each. Hence, the data analysis limitations of various forms of questionnaires should be considered before an instrument is designed. For example, less can be done statistically with open-ended questions than with ranking questions.

Chapter XIV: Interview Considerations

A. Overview

If properly used, the interview is an effective means of obtaining data. It is a technique in which an individual is questioned by a skilled and trained interviewer who records all replies, preferably verbatim in most cases. Most of the principals of questionnaire construction discussed in previous chapters pertain to the interview as well. This chapter, however, notes some issues specifically related to interviews.

Section XIV-B presents the distinction between structured and unstructured interviews. Interviewer's characteristics relative to the interviewee are noted in Section XIV-C. Situational factors are noted in Section XIV-D, while the topics of Sections XIV-E, F, and G are, respectively, training interviewers, data recording and reduction, and special problems. There is, unfortunately, little that can be recommended to avoid some of the problems noted in this chapter. The questionnaire developer should, in any case, be aware of them.

B. Structured and Unstructured Interviews

The term "structured" when applied to interviews is intended to emphasize that the interviewer employs a script of all the questions to be asked. In the unstructured interview, the interviewers may know many of the topics to be covered but they need to learn more about the subject overall, so they are willing to be led by the interviewee even into digressions. Unstructured interviews may occur as a preliminary to preparing either a questionnaire or a structured interview script. One could use a questionnaire as the script for a structured interview if one already had the questionnaire developed, but not enough time to convert it to a more convenient format. The main difference between the structured interview and questionnaire is procedural.

The degree of proficiency required of interviewers in conducting an unstructured interview is generally not available during Army field test evaluations. A structured interview requires the interviewer to have only moderate skill and proficiency, and hence is usually preferred. The advantages of the structured interview include: the opportunity to probe for all the facts when the respondent gives only a partial or incomplete response; a chance to ensure that the question is thoroughly understood by the respondent; and an opportunity to pursue other problem areas which may arise during an interview. The structured interview is almost always preferable to a questionnaire when the test group is small (10 to 20), and when time and test conditions permit.

As noted in Section II-B, unstructured interviews are not included within the definition of questionnaire used in this manual. They are, therefore, not discussed further.

C. Interviewer's Characteristics Relative to Interviewee

More research is needed to identify how characteristics of an interviewer affect the respondent. Some areas of concern are presented below.

1. Rank, Grade or Status of the Interviewer

For Army field test evaluations, it is recommended that the interviewer should be of similar rank or grade to the individuals being interviewed. A difference in rank or grade introduces a bias in the data which has been found to substantially influence test results. Interviewees tend to give the answer they perceive the higher-ranking interviewer favors. When the interviewer is of lower grade, the interviewee may not show respect and may not cooperate.

Evidence indicates that the greater the disparity between the status of the interviewer and that of the respondent, the greater the tendency for biased responses. Respondents tend to provide answers that will be more favorably received by the interviewer.

Data suggest that in the interview situation the respondent tends to support the norms adhered to by the interviewer. Lower socio-economic respondents may defer to the norms represented by a higher-status interviewer. The effect, however, is related to the types of questions asked. Sensitive issues involving socially accepted or rejected answers will effect more bias.

2. Sex of the Interviewer

Differences in response patterns according to the interviewer's sex depend on subject matter as well as on the composition of the respondent populations and other characteristics of the specific survey situation. Subject matter which tends to be most sensitive to differences in male/female response patterns deals with gender stereotypes. Interview items used in performance appraisals may be sensitive to sex role stereotypes. It is recommended that this type of item be investigated for rating differences between males and females. Interview items that are relevant to technical background experience (not usually obtained by females) also show gender response differences.

3. Race of the Interviewer

The effects of the race of the interviewer on the respondent should probably be viewed as the result of interaction between interviewer and respondent characteristics, or the result of the item content. Respondents often give socially-desirable answers to interviewers whose race differs from theirs, particularly if the interviewee's social status is lower than that of the interviewer and the topic of the question is threatening.

Nonsensitive, nonracial items appear to be relatively immune to interviewer effects for racial background. Therefore, racial background of the interviewer does not usually seem to affect survey results. It would be possible to assign interviewers of different racial background regardless of the respondent's racial background. An interviewer's race can probably establish different frames of reference for items with racially-related content. For threatening items or items with racially-related content, more valid results might be expected when the interviewer is of the same race as the respondent.

4. Experience of the Interviewer

There may be no significant differences between interview completion rates for experienced and inexperienced interviewers who have received sufficient interviewer training for face-to-face interviews and telephone interviews. However, it has been found that experienced interviewers may have different error rates than inexperienced interviewers. This error rate has been associated with the age of the interviewer, and the amount of interviewer training. Interviewer error is usually controlled through selection and training. Older interviewers (age 55 and over) have been known to frequently deviate from interviewer guides. Younger interviewers were found to follow the interview guides more closely. Nonstandardized administration of the interview could jeopardize the overall standardization of the survey procedures.

Other evidence indicates that field interviewers trained for less than a day produce more survey errors than more highly trained interviewers. Individuals responsible for developing interview items, guides, and training require sufficient development time prior to administration of the interview. Interview techniques to increase standardization have been known to improve through training. Response rates for telephone interviews may also be increased through training.

D. Situational Factors

Among the situational factors that should be considered when interviews are used are the following:

1. It helps greatly if the interviewees perceive the interviewer as interested in hearing their comments, as willing to listen, and (if the situation requires) as willing to protect them from recrimination for being adverse in their evaluations.

2. Interviews should be conducted in a quiet, temperature-controlled environment where the respondent can be comfortable and relaxed. Each respondent should be interviewed in private, separate and apart from all others, so that no other person hears or is biased by his/her responses.

3. The reinforcing behaviors of the interviewer have an influence on the responses collected, and at times may cause respondents to change their preferences. Such comments as "good" or "fine" and such actions as smiling and nodding can have a decided effect on test results. Praised respondents normally offer more answers than unpraised ones. Praising respondents may also tend to reduce "Don't know" answers without increasing insincere or dishonest responses.

4. Interested respondents seem to be more subject to interviewer effects than uninterested ones.

5. Interview questions which are read slowly indicate to respondents that they can take their time in carefully and thoughtfully answering the question. Rushing through an interview may reduce accuracy.

6. Use a "focus" group or pilot screening as a way to develop hypotheses and refine questions for establishing an interview guide and interview items. Interview guides are to be followed so that questions are asked without any wording changes. This promotes standardization across interviews.

7. Incomplete answers to survey questions require nondirective probing. When asking for clarification regarding an incomplete answer, the respondent is not to be directed toward any one response. Instead, phrases such as "tell me more" would be useful to employ.

8. Recording answers to interviews that use closed-end questions requires only that the interviewer mark the answer that the respondent selects.

9. When recording answers to open-ended questions, use a tape recorder if the respondent agrees or write down the answers verbatim. It is possible to combine open-ended and closed-end items for interview questionnaires, although coding and recording may be more difficult for the open-ended items.

10. For telephone surveys, use an interview structure and interview guide that promotes a high interaction between the interviewer and the respondent. This may be useful in increasing response rate.

11. Response cards can be adapted from face-to-face interviews 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.

E. Training Interviewers

Generally, interviewers require a certain amount of training. Army personnel may check with the Army Research Institute-Field Unit closest to them for help in this area. Some of the factors which should be considered when training interviewers are the following:

1. Training sessions for interviewers usually range between two days and five days. Interviewers conducting field interviews require more training than individuals who conduct telephone interviews. Two days minimum up through five days training are recommended for face-to-face interviews.

2. Sometimes researchers provide interviewers with information about the general research goals, sampling procedures, data analysis, and reports that will result from the survey.

3. Interviewer training requires general information in the course content such as how to introduce the study, as well as more specific information. Interviewers need to be familiar with the wording used in the survey, and any branching instructions. Standardization of the study through asking questions, probing incomplete answers, and recording answers are important aspects of the course content.

4. Interviewer training usually incorporates a demonstration of the standardized interview, and exercises where trainees role-play both the respondent and the interviewer. Practice sessions may also be tape recorded.

F. Data Recording and Reduction

In the structured interview, both questions and answers are orally communicated. The interviewer may encode the answers on paper, or tape record the responses for later encoding (but only if the interviewee agrees to the taping and does not seem influenced by the presence of a recording device).

Other topics related to interview data recording and reduction are outside the scope of this manual.

G. Special Interviewer Problems

This section notes some special problems related to interviews.

When interviews are used, the qualified interviewer will avoid leading, pressuring, or influencing the direction of an interviewee's evaluations. If potential interviewers have strong preferences regarding the system(s) being tested, they should probably be disqualified.

Many studies have been conducted that show other biasing effects on the interviewer. Factors leading to significant effects of the interviewer upon results include: relatively high ambiguity in the wording of the inquiry; interviewer "resistance" to a given question; and resistance to additional questioning or probing. Interviewer bias can exist without being apparent, and the direction of bias is not necessarily uniform. The least interviewer bias is probably found with questions that can be answered "Yes" or "No." The bias can result from differences in interviewing methods, differences in the degree of success in eliciting factual information, and differences in classifying the respondent's answers. Interviewers' expectations may have a more powerful effect on the results than their ideological preferences.

Some interviewers have a tendency not to transmit printed instructions word for word. Hence, total phrases may be eliminated and key words originally intended to focus the respondent's attention on some specific point are omitted or changed. Key ideas are lost, mainly through omission. Variability of interviewer performance seems to vary both across interviewers and within individuals.

An interviewer's attitude toward a question can communicate itself sufficiently to the respondent so that the meaning of the question is altered. When training interviewers to deliver a questionnaire in a standardized fashion, they need to rehearse the questions for tone of voice and body language to reduce any interviewer bias.

Research Product 89-21

Questionnaire Construction Manual Annex

Questionnaires: Literature Survey and Bibliography

June 1989

Fort Hood Field Unit
Systems Research Laboratory

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

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.

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

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

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

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

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

DD Form 1473, JUN 86 *Previous editions are obsolete.*

ARI Research Product 89-21

18. SUBJECT TERMS (Continued)

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

Research Product 89-21

Questionnaire Construction Manual Annex Questionnaires: Literature Survey and Bibliography

Bettina A. Babbitt
Essex Corp.

Charles O. Nystrom
U.S. Army Research Institute

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

Systems Research Laboratory
Robin L. Keesee, Director

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

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

June 1989

Army Project Number
2Q263739A793

Human Factors in Training and Operational Effectiveness

Approved for public release; distribution is unlimited.

FOREWORD

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

EDGAR M. JOHNSON
Technical Director

ACKNOWLEDGMENTS

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

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

QUESTIONNAIRE CONSTRUCTION MANUAL ANNEX
QUESTIONNAIRES: LITERATURE SURVEY AND BIBLIOGRAPHY

EXECUTIVE SUMMARY

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

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

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

QUESTIONNAIRE CONSTRUCTION MANUAL ANNEX
QUESTIONNAIRES: LITERATURE SURVEY AND BIBLIOGRAPHY

CONTENTS

		Page
I.	INTRODUCTION	1
II.	SCALE CATEGORIES	5
2.1	Multiple-Choice Scales	7
2.2	Bipolar Scales	23
2.3	Semantic Differential Scales	31
2.4	Rank Order Scales	37
2.5	Paired-Comparison Items	41
2.6	Continuous and Circular Scales	45
III.	BEHAVIORAL SCALES	51
3.1	Behaviorally Anchored Rating Scales	53
3.2	Behavioral Expectation Scales	59
3.3	Behavioral Observation Scales	63
3.4	Mixed Standard Scales	67
IV.	DESIGN OF QUESTIONNAIRE ITEMS	71
4.1	Open-Ended Items and Closed-End Items	73
4.2	Wording of Items and Tone of Wording	77
4.3	Length of Items and Number of Items	83
4.4	Order of Items	89
4.5	Balanced Items	93
V.	DESIGN OF SCALE CATEGORIES	99
5.1	Response Alternatives	101
5.2	"Don't Know" Category	109
5.3	Number of Scale Points	113
5.4	Middle Scale Point Position	121
VI.	INTERVIEWER AND RESPONDENT CHARACTERISTICS	127
6.1	Interviewing	129
6.2	Cognitive Complexity	137
6.3	Education	141
6.4	Ethnic Background	147
6.5	Gender	153
6.6	Age	159

CONTENTS (Continued)

		Page
VII.	QUESTIONNAIRE FORMAT	165
	7.1 Questionnaire Layout	167
	7.2 Branching	173
VIII.	FUTURE RESEARCH	177
BIBLIOGRAPHY		185
APPENDIX A.	P-77-2, QUESTIONNAIRE CONSTRUCTION MANUAL ANNEX. LITERATURE SURVEY AND BIBLIOGRAPHY: TABLE OF CONTENTS	225
B.	COMPARISON BETWEEN P-77-2, QUESTIONNAIRE CONSTRUCTION MANUAL ANNEX, AND THE SEQUEL	229
C.	OVERVIEW OF CONTENT AREAS COVERED BY P-77-2 AND THE SEQUEL	233
D.	FUTURE RESEARCH RECOMMENDATIONS	235

QUESTIONNAIRE CONSTRUCTION MANUAL ANNEX
QUESTIONNAIRES: LITERATURE SURVEY AND BIBLIOGRAPHY

CHAPTER I
INTRODUCTION

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

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

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

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

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

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

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

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

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

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

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

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

CHAPTER II

SCALE CATEGORIES

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

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

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

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

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

2.1 MULTIPLE-CHOICE SCALES

Description of Multiple-Choice Scales

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

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

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

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

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

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

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

Examples of Multiple-Choice Scales

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

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

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

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

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

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

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

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

(Select one descriptor)

___ "Demonstration by IP; no evaluation."

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

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

___ "Student required considerable verbal assistance."

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

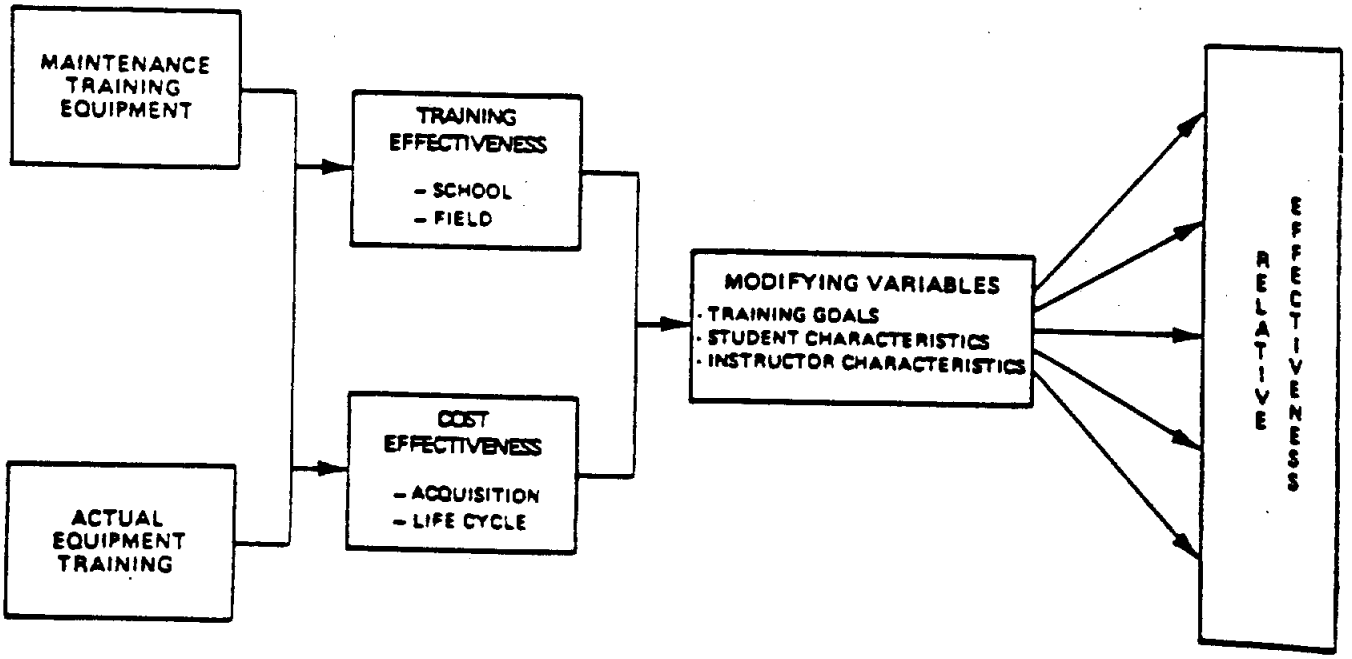
Interval item -- weighted. Equal distance between each scale point is assumed for interval scales. When constructing interval scales to measure the intensity of feeling, it is possible to design items where each response has a different weight assigned to it. The weights are used by analysts during analysis.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

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

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

*Weights not shown to respondents.

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



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

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

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

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

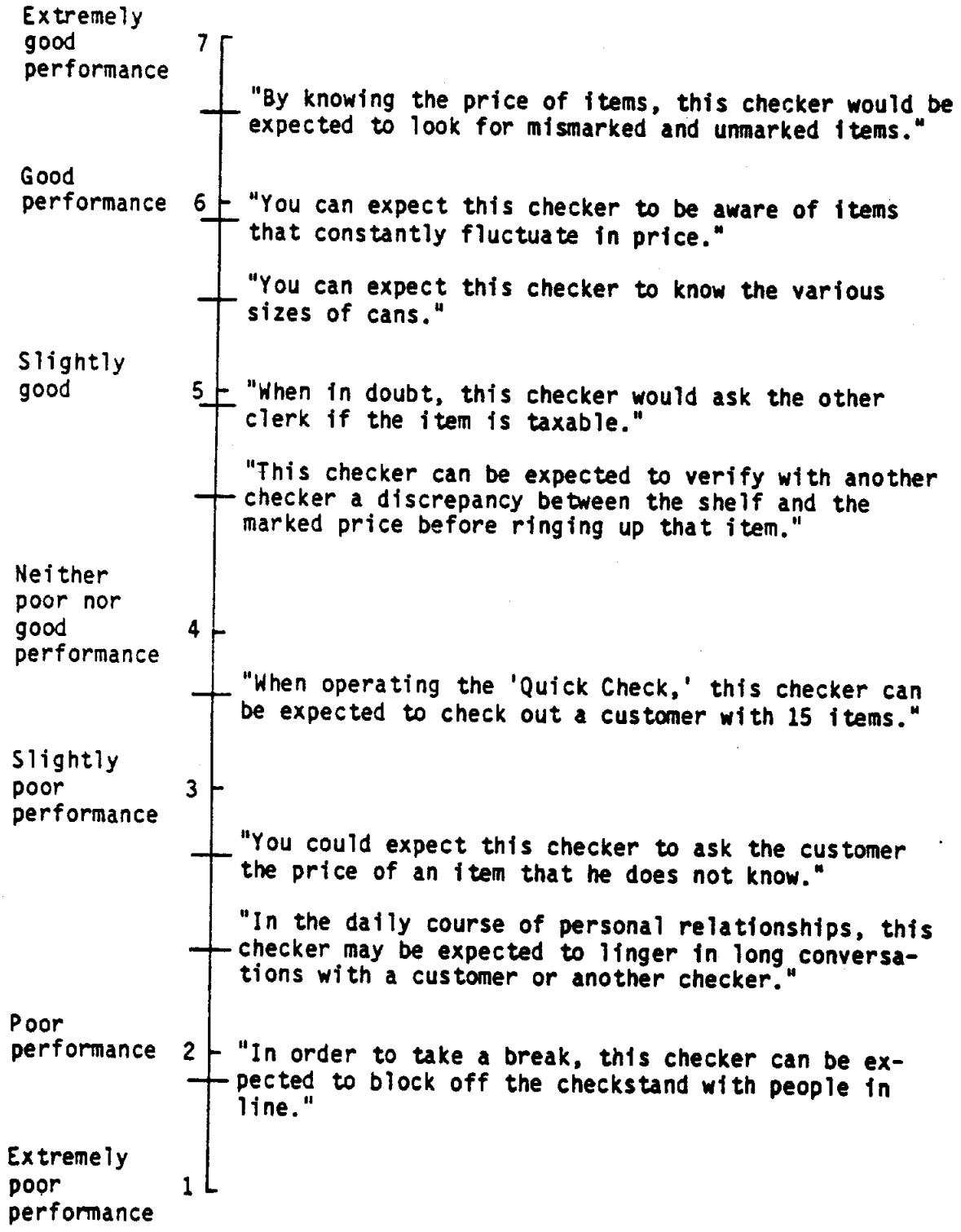
4. "System Understanding: Behaviors which show that the technician thoroughly understands system operation allowing him to recognize, diagnose, and correct problems not specifically covered in the Technical Orders and publications;"

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

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

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

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



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

Bipolar Nominally Contrasting Adjectives and Phrases

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

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

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

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

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

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

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

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

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

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

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

Agree ___ Disagree ___

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

Agree ___ Disagree ___

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

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

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

Hard to Agree

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

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

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

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

Easy to Agree

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

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

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

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

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

Comparisons of Multiple-Choice Scales

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

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

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

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

Conclusions Regarding Multiple-Choice Scales

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

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

2.2 BIPOLAR SCALES

Description of Bipolar Scales

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

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

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

Examples of Bipolar Scales

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

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

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

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

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

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

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

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

EFFECTIVENESS | INEFFECTIVENESS

+3 +2 +1 0 -1 -2 -3

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

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

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

Comparison of Bipolar Scales

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

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

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

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

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

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

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

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

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

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

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

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

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

Conclusions Regarding Bipolar Scales

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

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

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

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

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

2.3 SEMANTIC DIFFERENTIAL SCALES

Description of Semantic Differential Scales

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

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

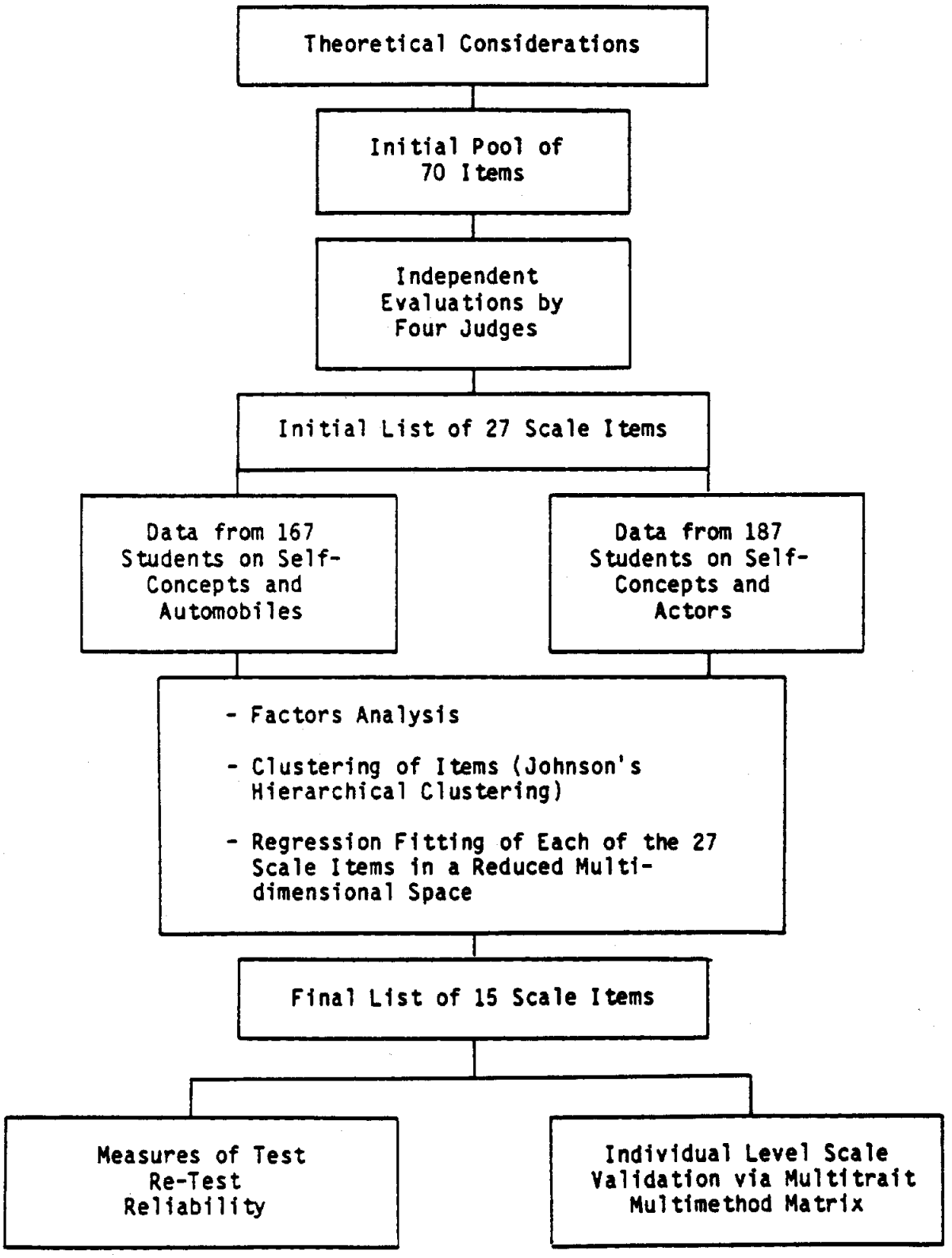
Semantic differential scales have been used by researchers in multiple fields, such as marketing, education, and psychotherapy. Marketing researchers have used this instrument extensively (Malhotra, 1981; Downs, 1978). These scales measure attitudes and opinions (Church, 1983). Attitudes may include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

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

Examples of Semantic Differential Scales

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

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



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

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

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

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

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

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

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

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

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

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

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

Comparisons of Semantic Differential Scales

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

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

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

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

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

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

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

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

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

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

Conclusions Regarding Semantic Differential Scales

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

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

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

Description of Rank Order Scales

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

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

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

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

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

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

Examples of Rank Order Scales

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

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

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

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

Comparisons of Rank Order Scales

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

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

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

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

Conclusions Regarding Rank Order Scales

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

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

2.5 PAIRED-COMPARISON ITEMS

Description of Paired-Comparison Items

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

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

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

Examples of Paired-Comparison Items

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

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

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

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

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

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

Comparisons of Paired-Comparison and Other Items

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

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

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

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

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

Conclusions Regarding Paired-Comparison Items

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

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

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

2.6 CONTINUOUS AND CIRCULAR SCALES

Description of Continuous and Circular Scales

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

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

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

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

Examples of Continuous and Circular Scales

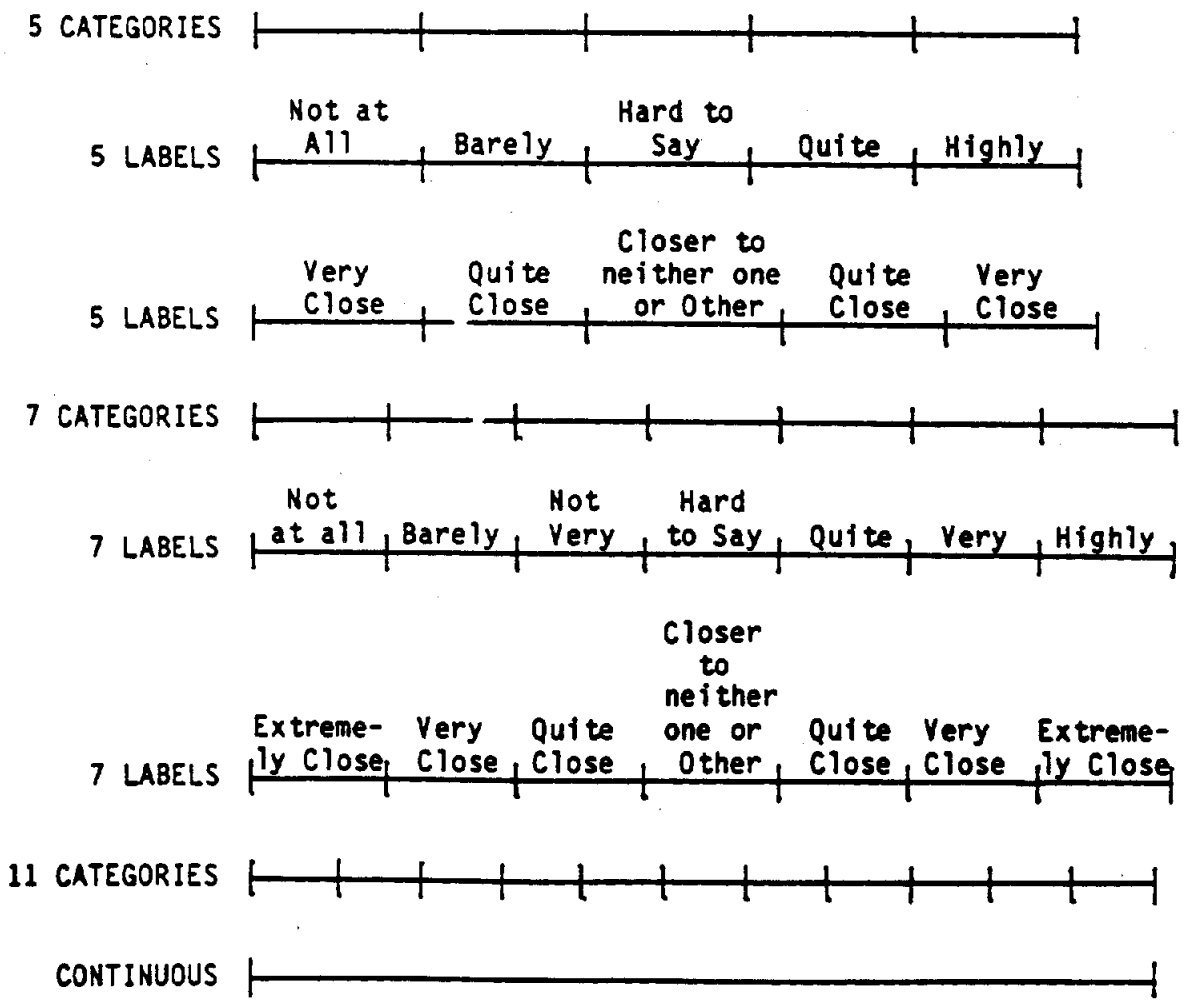
Continuous scales are usually thought of as straight lines with no indications of any differentiation along the scale lines. A continuous scale can provide the respondent with guidance as to the directionality of the rating, and offer the respondent greater discrimination as to ratings along the scale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

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

Friendly . . . . . Unfriendly

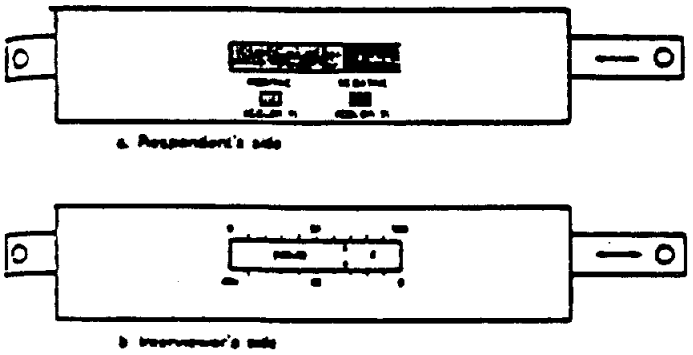
Friendly _____ Unfriendly

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



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

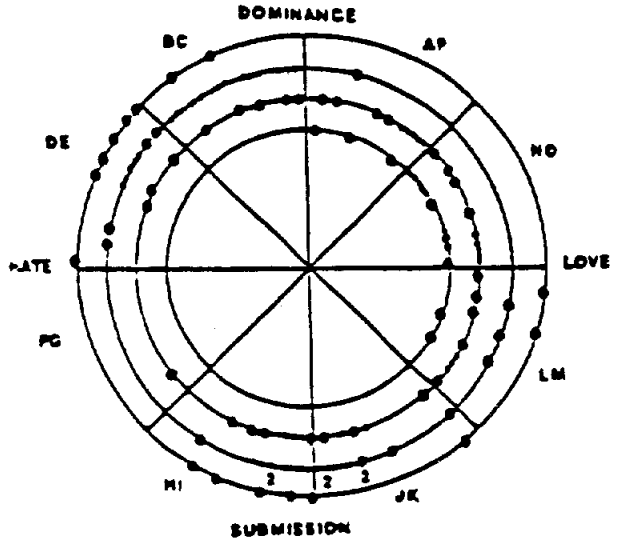
THE POLLIMETER (PATENT PENDING)



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

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

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



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

Comparisons of Continuous and Circular Scales

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

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

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

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

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

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

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

Conclusions Regarding Continuous and Circular Scales

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

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

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

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

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

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

CHAPTER III

BEHAVIORAL SCALES

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

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

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

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

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

3.1 BEHAVIORALLY ANCHORED RATING SCALES

Description of Behaviorally Anchored Rating Scales

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

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

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

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

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

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

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

Examples of Behaviorally Anchored Rating Scales

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

Scale Dimension

Performance and Effort on the Job

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

Comparisons of Behaviorally Anchored Rating Scales

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

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

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

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

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

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

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

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

Conclusions Regarding Behaviorally Anchored Rating Scales

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

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

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

3.2 BEHAVIORAL EXPECTATION SCALES

Description of Behavioral Expection Scales

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

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

Examples of Behavioral Expectation Scales

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

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

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

Very False 1 2 3 4 5 6 7 Very True

Comparisons of Behavioral Expectation Scales

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

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

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

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

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

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

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

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

Conclusions Regarding Behavioral Expectation Scales

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

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

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

3.3 BEHAVIORAL OBSERVATION SCALES

Description of Behavioral Observation Scales

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

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

Examples of Behavioral Observation Scales

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

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

Almost Never 1 2 3 4 5 Almost Always

Comparisons of Behavioral Observation Scales

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

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

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

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

Conclusions Regarding Behavioral Observation Scales

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

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

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

3.4 MIXED STANDARD SCALES

Description of Mixed Standard Scales

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

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

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

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

Examples of Mixed Standard Scales

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

MSS Error Counting System Anchors

Consistent Combinations:

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

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

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

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

Comparisons of Mixed Standard Scales

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

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

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

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

Conclusions Regarding Mixed Standard Scales

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

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

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

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

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

CHAPTER IV

DESIGN OF QUESTIONNAIRE ITEMS

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

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

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

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

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

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

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

4.1 OPEN-ENDED ITEMS AND CLOSED-END ITEMS

Description of Open-Ended Items and Closed-End Items

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

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

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

Examples of Open-Ended Items and Closed-End Items

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

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

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

_____

_____

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

_____

_____

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

_____

_____

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

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

Comparisons of Open-Ended Items and Closed-End Items

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

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

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

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

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

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

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

Conclusions Regarding Open-Ended Items and Closed-End Items

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

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

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

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

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

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

4.2 WORDING OF ITEMS AND TONE OF WORDING

Description of Wording of Items and Tone of Wording

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

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

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

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

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

Examples of Wording of Items and Tone of Wording

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

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

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

"Would you approve if the citizen . . ."

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

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

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

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

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

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

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

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

Comparisons of Wording of Items and Tone of Wording

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

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

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

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

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

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

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

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

Conclusions Regarding Wording of Items and Tone of Wording

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

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

4.3 LENGTH OF ITEMS AND NUMBER OF ITEMS

Description of Length of Items and Number of Items

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

Examples of Length of Items and Number of Items

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

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

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

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

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

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

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

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

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

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

Comparisons of Length of Items and Numbers of Items

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

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

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

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

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

Conclusions Regarding Length of Items and Number of Items

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

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

4.4 ORDER OF ITEMS

Description of Order of Items

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

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

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

Examples of Order of Items

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

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

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

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

Comparisons of Order of Items

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

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

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

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

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

Conclusions Regarding Order of Items

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

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

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

4.5 BALANCED ITEMS

Description of Balanced Items

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

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

Examples of Balanced Items

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

Hot ____ Cool

Hot ___X___ Cold

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

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

Item

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

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

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

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

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

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

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

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

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

Comparisons of Balanced Items

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

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

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

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

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

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

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

Conclusions Regarding Balanced Items

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

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

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

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

CHAPTER V

DESIGN OF SCALE CATEGORIES

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

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

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

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

Description of Response Alternatives

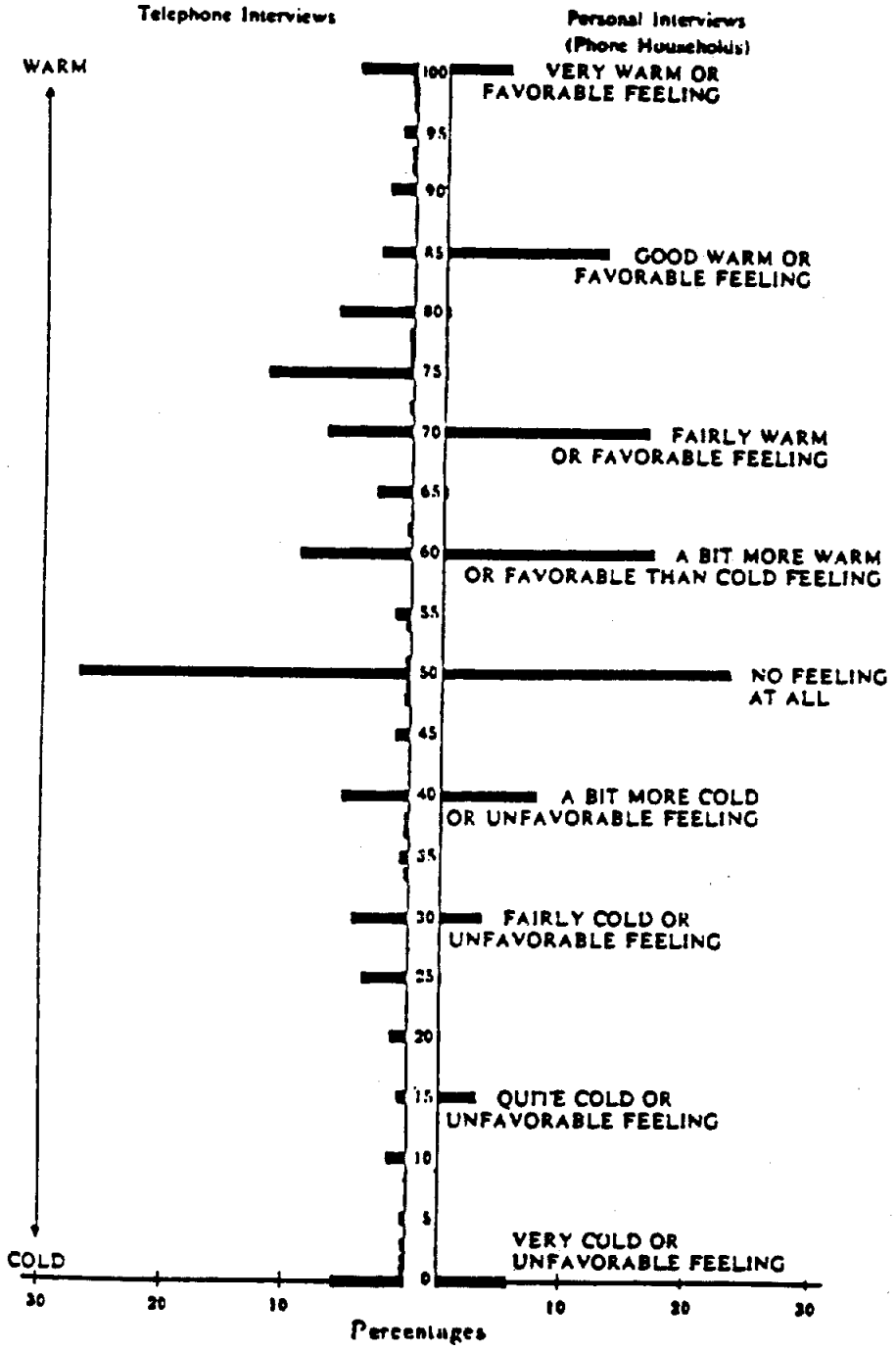
Points along the continuum of a scale have been identified (anchored) by numbers, adjectives, adverbs, description of behaviors, simple words, phrases, and complete sentences. Even the frequency and pattern of assigning anchors to the scale points has been varied. Some scales are completely anchored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

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

Examples of Response Alternatives

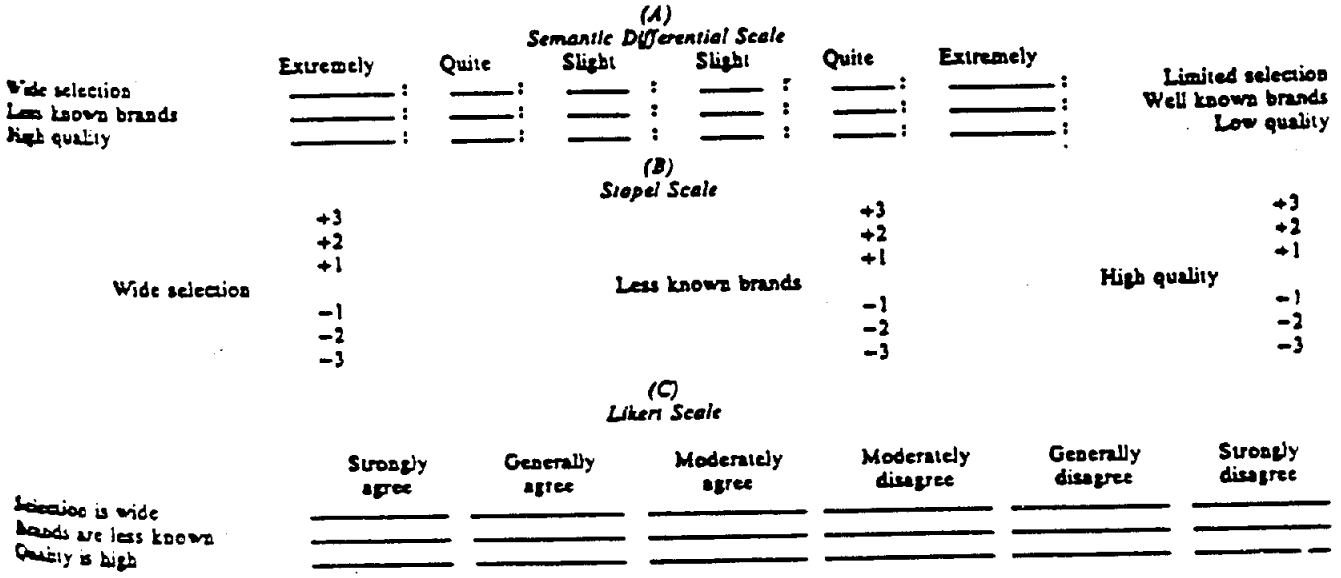
Backstrom and Hurchur-Cesar (1981) developed anchors for items they considered sensitive or that they felt required complicated responses. They used cards having precoded responses printed on them for sensitive items, and also for a lengthy series of questions requiring complicated responses. Use of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

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



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

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



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

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

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

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

RESULTS PERTAINING TO 'ACCEPTABILITY' DESCRIPTORS

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

RESULTS PERTAINING TO 'ACCEPTABILITY' DESCRIPTORS (Cont.)

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

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

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

Comparisons of Response Alternatives

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

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

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

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

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

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

Conclusions Regarding Response Alternatives

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

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

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

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

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

5.2 "DON'T KNOW" CATEGORY

Description of the "Don't Know" Category

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

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

Examples of the "Don't Know" Category

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

Schuman and Presser "Don't Know" Filter Experiments

1. Courts (NORC-74)

Standard Form

"In general, do you think the courts in this area deal too harshly or not harshly enough with criminals?"

Too harshly	5.6%
Not harshly enough	77.8%
About right (volunteered)	9.7%
Don't know (volunteered)	6.8%
	(N=745)

Filtered Form

"In general, do you think the courts in this area deal too harshly or not harshly enough with criminals, or don't you have enough information about the courts to say?"

Too harshly	4.6%
Not harshly enough	60.3%
About right (volunteered)	6.1%
Not enough information to say	29.0%
	(N=723)

2. Government (SRC-76 February)

"Some people are afraid the government in Washington is getting too powerful for the good of the country and the individual person. Others feel that the government in Washington is not getting too strong. What is your feeling, do you think the government is getting too powerful or do you think the government is not getting too strong?"

Too powerful	55.0%
Not too strong	35.1%
Don't know (volunteered)	10.0%
	(N=613)

"Some people are afraid the government in Washington is getting too powerful for the good of the country and the individual person. Others feel that the government in Washington is not getting too strong. Have you been interested enough in this to favor one side over the other? (If yes) What is your feeling, do you think the government is getting too powerful or do you think the government is not getting too strong?"

Too powerful	45.0%
Not too strong	21.6%
Not interested enough	33.3%
	(N=606)

3. Communist Book (SRC-77 February)

"This next question is about a man who admits he is a communist. Suppose he wrote a book which is in your public library. Somebody in your community suggests the book should be removed from the library. Would you favor removing the book or oppose removing the book?"

Favor removing	29.1%
Oppose removing	67.9%
Don't know (volunteered)	3.0%
	(N=563)

"This next question is about a man who admits he is a communist. Suppose he wrote a book which is in your public library. Somebody in your community suggests the book should be removed from the library. Would you favor removing the book, oppose removing the book, or do you not have an opinion on that?"

Favor removing	17.2%
Oppose removing	56.6%
No opinion	26.2%
	(N=533)

Comparisons of "Don't Know" Categories

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

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

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

Conclusions Regarding the "Don't Know" Category

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

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

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

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

5.3 NUMBER OF SCALE POINTS

Description of Number of Scale Points

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

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

Examples of Number of Scale Points

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

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

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

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

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

Nystrom Number Scale

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

Maintenance Vehicle Questionnaire by Krohn

Ease of Use Rating Scale

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

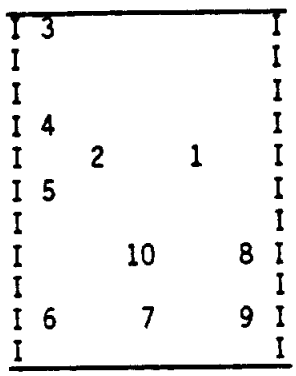
How easily can you:

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

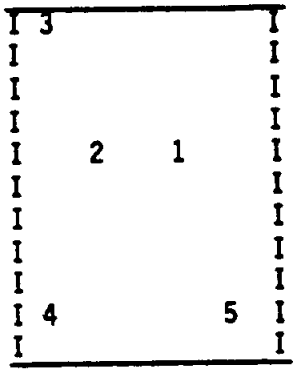
The Nystrom Number Scale includes directions for the respondents. An illustration for respondent directions is presented here for the operation under usual conditions, Bradley Fighting Vehicle Test:

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

IFV



CFV



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

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

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

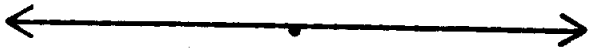
INEFFECTIVE | EFFECTIVE
-3 -2 -1 0 +1 +2 +3

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

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

Basic Rating Scales

Attribute



negative neutral positive

Bipolar, 5-interval, Horizontal Scale

good - - - - - bad

Unipolar, 10-interval, Vertical Scale

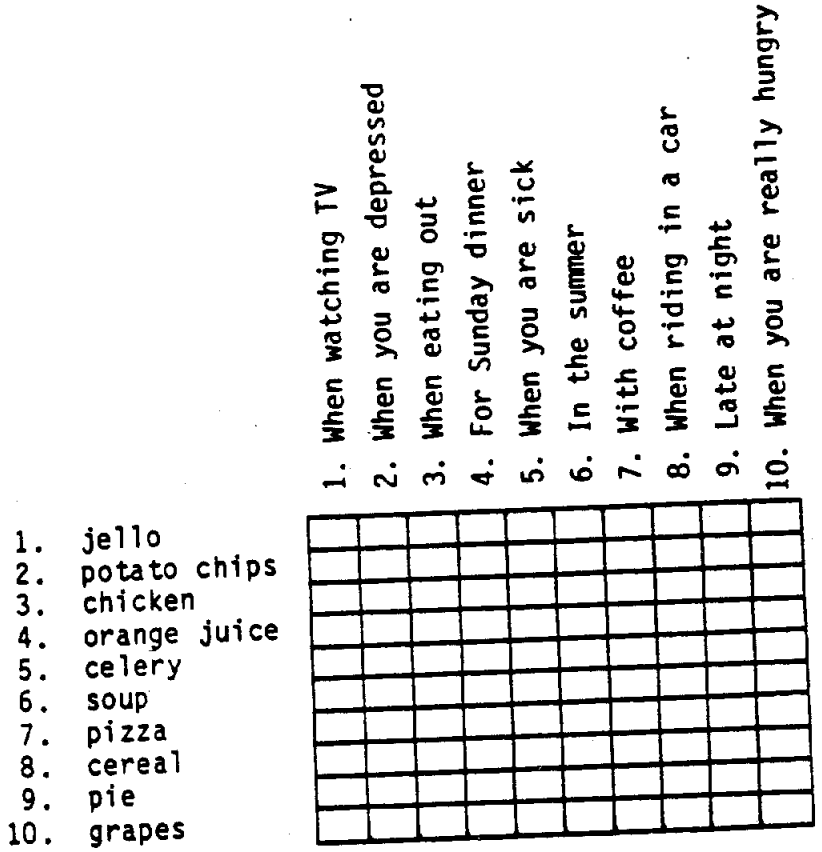
Humorousness

+5 —
+4 —
+3 —
+2 —
+1 —
-1 —
-2 —
-3 —
-4 —
-5 —

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

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

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



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

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

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

Comparisons of Number of Scale Points

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

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

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

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

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

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

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

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

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

Conclusions Regarding Number of Scale Points

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

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

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

5.4 MIDDLE SCALE POINT POSITION

Description of Middle Scale Point Position

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

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

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

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

Examples of Middle Scale Point Position

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

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

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

> Do you feel that the state government has too much or too little control over local law enforcement training?
>
> 1. Too Much 5. Too Little 3. (If Volunteered) Right Amount
>
> Do you feel that the state government has too much, too little, or the right amount of control over local law enforcement training?
>
> 1. Too Much 5. Too Little 3. Right Amount

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

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

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

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

Comparisons of Middle Scale Point Position

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

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

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

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

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

Conclusions Regarding Middle Scale Point Position

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

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

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

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

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

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

CHAPTER VI

INTERVIEWER AND RESPONDENT CHARACTERISTICS

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

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

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

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

6.1 INTERVIEWING

Description of Interviewing

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

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

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

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

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

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

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

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

Examples of Interviewing

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

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

SPECIFIC SCENARIO INTERVIEW OUTLINE

TASK: EXTRACT, REPLACE AND TRANSPORT EQUIPMENT

EQUIPMENT PROBLEMS

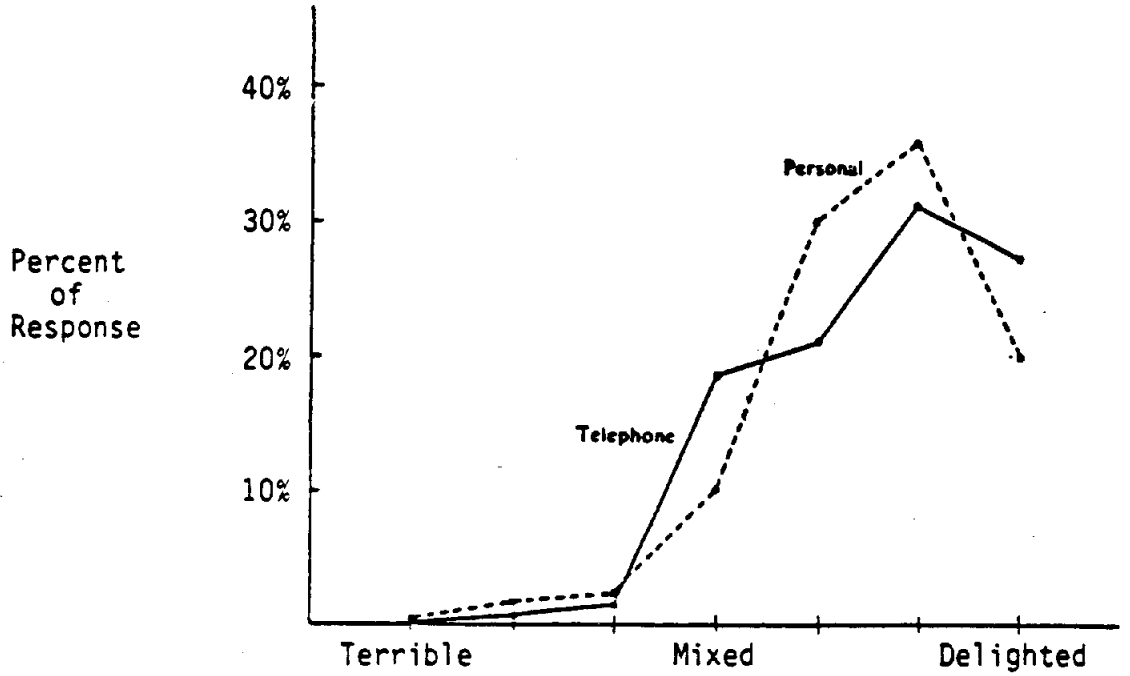
"I will name equipment from the LAVM/RV that you may have used to extract, replace and transport equipment. Please answer Yes or No to indicate whether or not you experienced any difficulties using the equipment. I would also appreciate your comments concerning the difficulties. If you have no experience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

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

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

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

Telephone and Personal Interviews
Delighted-Terrible Scale Response Distribution



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

Comparisons of Interviewing

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

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

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

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

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

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

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

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

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

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

Conclusions Regarding Interviewing

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

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

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

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

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

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

6.2 COGNITIVE COMPLEXITY

Description of Cognitive Complexity

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

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

Examples of Cognitive Complexity

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

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

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

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

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

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

Comparisons of Cognitive Complexity

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

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

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

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

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

Conclusions Regarding Cognitive Complexity

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

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

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

6.3 EDUCATION

Description of Education

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

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

Examples of Education

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

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

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

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

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

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

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

Comparisons of Education

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

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

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

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

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

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

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

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

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

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

Conclusions Regarding Education

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

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

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

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

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

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

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

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

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

6.4 ETHNIC BACKGROUND

Description of Ethnic Background

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

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

Examples of Ethnic Background

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

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

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

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

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

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

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

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

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

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

Comparisons of Ethnic Background

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

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

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

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

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

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

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

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

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

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

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

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

Conclusions Regarding Ethnic Background

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

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

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

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

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

Description of Gender

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

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

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

Examples of Gender

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

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

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

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

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

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

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

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

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

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

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

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

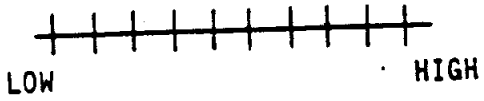
   a. Very interested
   b. Somewhat interested
   c. Not very interested

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

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

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

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



Comparisons of Gender

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

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

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

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

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

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

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

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

Conclusions Regarding Gender

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

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

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

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

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

6.6 AGE

Description of Age

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

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

Examples of Age

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

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

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

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

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

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

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

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

Comparisons of Age

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

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

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

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

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

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

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

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

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

Conclusions Regarding Age

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

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

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

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

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

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

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

CHAPTER VII

QUESTIONNAIRE FORMAT

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

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

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

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

7.1 QUESTIONNAIRE LAYOUT

Description of Questionnaire Layout

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

Examples of Questionnaire Layout

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

Example of the Unstructured Observation Form

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

   "What errors did you observe?" _____

   _____

   _____

Example of the Semistructured Observation Form

PROBLEM 1 : AMPLITUDE MEASUREMENT

A. PRELIMINARY ADJUSTMENTS

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

MAXIMUM POINTS (4.0)
POINTS ASSIGNED: _____

B. CONTROL SETTINGS

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

MAXIMUM POINTS (4.0)
POINTS ASSIGNED: _____

C. WAVEFORM ANALYSIS

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

MAXIMUM POINTS (15.0)
POINTS ASSIGNED: _____

D. SAFETY

MAXIMUM POINTS (2.0)
POINTS ASSIGNED: _____

PROBLEM TOTAL _____

PASSED ☐ FAILED ☐

Example of the Structured Observation Form

PROBLEM 1 : AMPLITUDE MEASUREMENT

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

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

PASSED ☐ FAILED ☐

ORIGINAL FEED DIRECTION

THIS SIDE UP FOR

BLANK

INSERTS

[illegible]

[illegible] (use reverse side [illegible]

[illegible] DOCUMENT [illegible]

[illegible] sheet [illegible]

- EACH INSERT MUST BE FOLL[illegible]
- KEY SHEETS WILL NOT BE COPIED [illegible]
- SELECT *INSERTS*—YES ON SPECIAL [illegible]

A-2764

Property of
DEFENSE TECHNICAL INFORMATION CENTER

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN DIEGO CALIFORNIA 92152

NPRDC SR 78-8

APRIL 1978

HUMAN FACTORS IN OPERATIONAL SYSTEM TESTING: A MANUAL OF PROCEDURES

Special Report 78-8 April 1978

HUMAN FACTORS IN OPERATIONAL SYSTEM TESTING: A MANUAL OF PROCEDURES

D. Meister

Reviewed by
F. Muckler

Approved by
James J. Regan
Technical Director

Navy Personnel Research and Development Center
San Diego, California 92152

FOREWORD

This manual was developed under Exploratory Development Work Unit Number 521.101.03.10: U. S. Marine Corps Human Factors Analysis. It describes procedures for personnel performance evaluation in the context of an Operational System Test (OST).

Although these procedures have been developed specifically for the U. S. Marine Corps, it is believed that they are applicable to personnel performance evaluation in all military services.

The cooperation of the U. S. Marine Corps Headquarters (MPI-20) and of the Marine Corps Development and Education Command (OT&E), Quantico, Virginia is much appreciated.

Section EIGHT: Introduction to Statistical Methodology was authored by Dr. Jules Borack.

J. J. CLARKIN
Commanding Officer

CONTENTS

	Page
SECTION ONE--INTRODUCTION	1-1
SECTION TWO--THE PERSONNEL PERFORMANCE TEST PLAN	2-1
Introduction	2-1
The Purpose of Personnel Performance Testing	2-1
Outline of Personnel Performance Test Plan	2-4
Example of Personnel Performance Test Plan--Personnel Performance Measurement in the XM-47 Amphibious Troop Transport and Tank Destroyer	2-16
Development of Quantitative Performance Criteria	2-28
SECTION THREE--THE SELECTION AND DEVELOPMENT OF MEASURES AND MEASUREMENT METHODS	3-1
Introduction	3-1
Objective Measures and Methods	3-11
Subjective Measures and Methods	3-31
Summary	3-56
SECTION FOUR--HUMAN ENGINEERING CHECKLIST PROCEDURES	4-1
Introduction	4-1
Human Engineering Checklist	4-3
SECTION FIVE--SELF REPORT RATING SCALES	5-1
Introduction	5-1
Rating Scales	5-1
SECTION SIX--INTERVIEW QUESTIONS	6-1
SECTION SEVEN--TEST PROCEDURES	7-1
SECTION EIGHT--INTRODUCTION TO STATISTICAL METHODOLOGY	8-1
Introduction	8-1
Statement of the Problem	8-1
Variables	8-2
Descriptive Statistics	8-4
Statistical Inference	8-12
Sampling Techniques	8-15
Summary	8-16
References	8-18
SECTION NINE--PERSONNEL PERFORMANCE TEST PLANNER'S CHECKLIST	9-1

Page

SECTION TEN--THE PERSONNEL PERFORMANCE TEST REPORT 10-1

Introduction 10-1
Outline of the Personnel Performance Test Report 10-1

SECTION ELEVEN--USEFUL REFERENCES 11-1

Specifications and Standards 11-1
Reference Books 11-1

SECTION TWELVE--INDEX 12-1

DISTRIBUTION LIST

SECTION ONE--INTRODUCTION

In accordance with MCO P5000.11, Test and Evaluation of Systems and Equipment for Operating Forces of the Marine Corps, the U. S. Marine Corps performs Operational System Tests to determine that new man-machine systems entering the Marine Corps inventory will satisfy Corps requirements.

Such systems consist of equipment; personnel who will operate and maintain the equipment; and the procedures, logistics, technical documentation, etc. necessary to support the personnel. The evaluation of each new system therefore must include a determination that personnel can in fact efficiently operate and maintain that system in accordance with Marine Corps requirements.

Experience with many systems acquired not only by the Marine Corps but also the other Services indicate that, if personnel have difficulty performing their taskvis, ithe new system, no matter how sophisticated and well engineered, will not function as effectively as it should. Each equipment and system imposes its own burdens on the personnel operating and maintaining that system. To use a very simplified example, if a rifle is excessively heavy, it will take longer for the rifleman to aim the rifle and his rate of fire will thereby be reduced. The more complex the new system, the more likely that it will contain features that increase the difficulty of using it.

Despite this, military testing of new systems often tends to ignore the personnel performance aspect. Many testers feel that the human is so flexible that he can adapt to and overcome equipment characteristics that create difficulties. Ordinarily this is correct, but combat--in which the new system may eventually be tested--is not an ordinary situation and severely stresses the Marine. Under stress conditions, minor equipment deficiencies that could otherwise be lived with become more difficult to overcome. Under these circumstances the human is not as flexible and adaptive as he is ordinarily.

Those authorized to evaluate newly acquired systems have a significant responsibility for uncovering minor equipment deficiencies. They are in a key position to improve the quality of Marine Corps equipment by determining where inadequacies exist and by recommending ways of eliminating them.

The title of this manual is "Human Factors in Operational System Testing: A Manual of Procedures." Human Factors refers to the entire complex of elements that affect personnel performance. These elements include (1) the way in which the equipment has been designed (equipment characteristics), (2) the procedures developed to operate and maintain the equipment, (3) the training provided to enable effective operation and maintenance, (4) the performance aids supplied to assist operators and maintainers (e.g., technical manuals), and (5) the environment in which the equipment will be used. The discipline of Human Factors has, on the basis of research conducted and experience gained over the 35 years since World War II, established personnel standards for each of the above elements.

A major part of Human Factors work is the development of procedures for testing to determine whether new equipment and systems meet these personnel standards. The procedures in this manual have been developed to permit Marine Corps test personnel to evaluate the new system for its human factors adequacy. They have been specially designed to require as little specialized Human Factors knowledge as possible. Where significant personnel-related problems arise, however, a Human Factors specialist should, of course, be consulted.

No one can force the test planner and test conductor to include these procedures in his test operations. If they are ignored, however, the resultant test and the conclusions derived from the test will be inadequate to answer the basic question of whether the system is one that the Marine Corps wants in its inventory. Further, the chances are excellent that the system, if accepted, will have degraded personnel capabilities.

SECTION TWO--THE PERSONNEL PERFORMANCE TEST PLAN

This section provides information on the Personnel Performance Test Plan, including its purpose, a step-by-step description of its various sections and requirements for completing those sections, a model to illustrate the material to be included in such a plan, and a procedure for developing quantitative personnel performance criteria.

Introduction

Operational System Testing (OST) is the process by which a military service verifies that a newly acquired system can in fact perform in accordance with military requirements. OST is required of the U. S. Marine Corps in its Systems Acquisition Management Manual (MCO P5000.10). An essential part of that verification process is the determination that system personnel--those who will operate and maintain the system--can perform required tasks to a required level of capability.

The first and essential step in the measurement of personnel performance is to develop a Personnel Performance Test Plan. The purposes of this section are to describe, in simple, step-by-step fashion, how to develop that test plan and the major points to be covered, and to illustrate it with several examples. This test plan must, of course, satisfy the format established in MCO P5000.11, Test and Evaluation of System and Equipment For Operating Forces of the U. S. Marine Corps.

Development of a Personnel Performance Test Plan is only the first of the activities required in measuring personnel performance during OST. It is, however, critical to the planning of a number of additional activities, including the following:

1. Development of performance criteria.

2. Selection of performance measures.

3. Selection of subjects (the system operators and maintainers whose performance is to be measured).

4. Determination of methods of measurement.

5. Methods of analyzing and reporting performance data.

The test plan is the basic document describing how the personnel part of the system will be tested.

The Purpose of Personnel Performance Testing

Before a new system can be accepted into inventory by a military service, it is necessary to verify that the system can perform its mission to specified requirements. Verification is a short-hand phrase which includes the following test goals:

1. To assess the accomplishment of system development objectives.

2. To ensure that systems, equipment, and personnel meet established requirements.

3. To forecast how the system will perform in an actual operational situation.

4. To ensure the effective integration of all elements of the system (including equipment, personnel, technical data, supplies, etc.).

5. To detect operational and engineering deficiencies in time for changes to be incorporated before significant production build-up.

6. To provide data and operational analyses for application to current and future systems.

7. To identify manpower and personnel resources needed to support the operational system.

8. To provide information to validate forthcoming training programs.

All of the above involve personnel testing. OST assumes that:

1. The system being evaluated is a prototype of the operational system and is configured as much as possible as it will be in its eventual operational use. If the system being tested deviates significantly from its operational configuration, the conclusions derived from testing will not describe what the system will be able to do operationally. What this means in effect is that personnel operating and maintaining the equipment during OST should be required to perform as much as possible as they would once the system is put into actual use. In other words: Don't take short cuts that would not be allowed in actual operation; don't give test personnel assistance they would not ordinarily have; and carry the test out to its ultimate conclusion and don't abort because something goes wrong (unless it imperils the system and its personnel). If the system is not operated according the operational procedures, conclusions cannot be drawn about how that system will function in the real world. Of course, it is impossible to fully simulate a combat situation, but test conditions should approximate this as much as possible.

2. Personnel operating and maintaining the system being evaluated should have characteristics similar to those that operational personnel will have and are required to perform as they would operationally. If the test conductor has a choice, he should select test personnel who are "average" in their past performance, not those whom he considers to be most proficient. Test personnel who, as a group, have a range of capability are most desirable. The most proficient personnel will make the system appear to be more efficient than it will actually turn out to be when it must be operated by average Marines. If there are any weak points in the system, it will be easier to find these with average personnel than with those who are much more proficient; the latter can compensate for these weak points, the former cannot.

3. All aspects of the system (i.e., equipment, personnel, logistics, procedures, and data) should be measured as an integral part of the evaluation. How the system functions is determined by all its elements. Ignoring one or more of these elements means that the system will not function realistically. Moreover, one cannot extract a system element and test it apart from other elements.

4. Once OST has begun, let the test personnel assume control of system operations; that is, the system should be allowed to perform without interference from Test Management (except to avoid conditions that may affect personnel safety or the integrity of the system).

The term "system" as used henceforth in this section includes not only an entire system, but any subsystem or equipment unit of the system (together with its operating personnel).

Logic requires that OST involve measurement of personnel performance. Every system procured for a military service includes not only equipment, but also operating and maintaining personnel, together with the procedures for these operations. Personnel form a subsystem of the total system; and the total system cannot perform its mission adequately unless <u>all</u> its subsystems function in accordance with requirements. The system will therefore not function properly unless its personnel also function properly. System adequacy cannot therefore be verified unless it is certain that system personnel can operate and maintain the system to specified standards.

Tests of engineering capability designed to measure physical characteristics of the system (e.g., fuel consumption, rated engine power, missile penetrating force) do not automatically measure personnel performance. Consequently the OST must make special provisions for measurement of personnel performance. Unless this is done, the OST will lack validity. The overestimation of system capability by developers often results from failure to factor in the ability of personnel to utilize the system. Since personnel performance often degrades system effectiveness, lack of personnel performance measurement data requires the system evaluator to assume perfect performance on the part of personnel--which obviously is not often true.

It is often assumed by those unfamiliar with the process that personnel performance measurement is an expense added to the already burdensome cost of OST. However, since personnel performance is inherent in system operations, and since the system must be exercised as it would be in routine operations, personnel performance measurement can be coordinated with other on-going aspects of the OST. It usually requires no additional test time beyond that required for on-going, already scheduled tests. There may be occasions when a special question must be examined and the overall OST does not contain provisions for examining the question; a special test may then be required. The number of such special tests will, however, be minimal.

In most cases, personnel performance measurement also requires no, or only minor, special instrumentation. It does, however, require the assignment of a limited number of personnel to collect and analyze the necessary data. The cost of this personnel time is relatively slight considering the importance of securing a quantitative measure of personnel effectiveness.

The measurement process does, however, require special test procedures that are described in the Personnel Performance Test Plan. This plan is not the overall system test plan but merely one part of it; most often it will be incorporated in the test plan as a special annex.

Development of the test plan is not merely a pencil-and-paper exercise. It is needed for the following reasons. First, writing the plan ensures that those who are responsible for performance measurement and others involved in the process (like data collectors) know exactly what they are supposed to do. Failure to write the plan often means that essential measurement elements are overlooked. Second, Test Management personnel often require that such a plan be written as a preliminary to conducting the OST. This is necessary if they are to know what is going on.

Outline of Personnel Performance Test Plan

Figure 2-1 provides an abbreviated outline of a Personnel Performance Test Plan (a full-scale test plan would contain much more detail). The examples provided under the various sections of the plan do not refer to any existing U. S. Marine system or equipment. They may imply a rather complete field test, which would be appropriate for the evaluation of a major system like a tank, an aircraft, or a fire direction center. The size and complexity of the system being evaluated (including the functions to be performed by personnel) will usually determine the scope of the testing. If a subsystem or a unit of a system is under evaluation (e.g., a single console), it may not be necessary to exercise the entire system of which the item under test is only an element. For example, if one were evaluating a new rifle, it would not be necessary to engage in regimental maneuvers to test that rifle. The scale of testing is determined by the number and type of inputs needed to reproduce the operational situation in which the system, subsystem, or unit will be used. In the case of the rifle, for example, this would probably be only a firing range reproducing the types of targets and the terrain in which one would use the rifle.

The same qualification applies when the purpose of the test--the specific question to be answered--is highly specialized. Assume that the evaluator wished to check to make sure that the interior environment of a new tank--temperature, humidity, air flow--met the standards of MIS STD 1472B (Department of Defense, 1974),[1] which is the governing document for the human engineering characteristics of man-machine systems. If that were the only question to be answered by the test, it might not be necessary to exercise the tank in simulated combat in order to secure an answer. Driving the tank for specified periods of time during the day and night might provide the needed data. As in the case of a unit of a system, the scale of testing required to answer a particular question is determined by the number and type of inputs needed to provide a reasonable simulation of the operational situation.

---

[1]Department of Defense. MIL STD 1472B. Military Standard, Human Engineering Design Criteria for Military Systems, Equipment, and Facilities, 31 December 1974.

1.0 PURPOSE

1.1 General. Example: Verify that system personnel can perform required tasks.

1.2 Specific. Example: Determine the type and magnitude of errors made by personnel; determine the effect of low temperature (Arctic) conditions on personnel ability to maintain a tank.

2.0 DESCRIPTION OF SYSTEM BEING EVALUATED

2.1 List of equipments to be operated (maintained) by personnel and for which personnel performance data are to be collected.

Example: _____ gun mount; _____ grenade launcher.

2.2 List of equipment tests during which personnel performance data are to be collected.

Example: Installation and checkout of the _____ fire direction console.

2.3 List of tasks for which personnel performance data will be collected.

Example: Alignment of the _____ theodolite.

2.4 Applicable Technical Manuals or other procedures.

Example: TM _____, Operation of the _____ machine gun;

TM _____, Operation of the _____ landing vehicle, personnel.

3.0 SPECIAL COMPARISONS (optional)

Example: Comparison of daytime and nighttime reconnaissance in jungle terrain.

4.0 CRITERIA AND MEASURES

4.1 Personnel performance criteria. Example: Receive, code, and transmit between 12 and 15 messages per hour; fire minimum of 5 rounds within 7 minutes.

4.2 Personnel performance measures. Example: Time taken to load hand-held missile; officer evaluation of squad reconnaissance performance.

Figure 2-1. Outline of the Personnel Performance Test Plan.

5.0 DATA COLLECTION METHODS

5.1 Data collectors

5.1.1 Number. Example: 4

5.1.2 Tasks to be performed. Example: Record start/stop time in operation of laser tracking set.

5.1.3 Training (if required). Example: All data collectors will receive 3 hours instruction in gathering data on the retractable machine gun (see training schedule appended).

5.2 Data collection forms. Example: Data sheet for retractable machine gun (appended); Post mission debriefing questionnaire for tank driver (appended).

5.3 Data collection procedures. Example: See Appendix A.

5.4 Instrumentation (only if required). Example: 2 tape recorders with 3 rolls of tape per data collector.

6.0 SUBJECTS

6.1 Number. Example: 3 squads.

6.2 Required characteristics. Example: ALL subjects will have 20/20 vision (corrected) and will have been qualified in operation of the retractable machine gun.

7.0 DATA ANALYSIS

Example: Determine mean number (and standard deviation) of messages transmitted between forward observers and batteries; develop equation relating gun loading speed and operator errors.

8.0 TESTING SCHEDULE

Example: Concurrent with other tests.

Figure 2-1 (Continued).

The principles described in this section apply regardless of the size of the unit being tested or the scope of the question being asked. In all cases, it is necessary to develop a Personnel Performance Test Plan, including the analyses involved in developing that plan. If the size of the unit to be tested is small, or the questions to be answered are few, the scope of the test plan can be scaled down. However, all sections of the test plan must be completed.

The following paragraphs describe the various sections of the test plan. The reader should refer to Figure 2-1 as each section is described.

Section 1.0--PURPOSE

This section describes the purpose of the Personnel Performance Test. The test has two purposes, general and specific. Of the two, the latter is by far the more important.

Although the purpose of personnel performance measurement may seem obvious at first glance, it is not. Thus, that purpose must be specified in detail.

The general purpose is to verify that personnel can perform their assigned tasks within the new system to meet system requirements and to ensure that personnel functions are effectively integrated with other system elements (see preceding section entitled "Purpose of Personnel Performance Testing"). The general purpose points to the necessity for specifying (preferably in quantitative form) the standards that the system and the personnel must meet. It emphasizes the necessity of comparing actual and desired personnel performance.

Unfortunately, the general purpose does not indicate what and how one should measure. Although it is important to emphasize that the general purpose is to verify personnel performance adequacy, this purpose is not specific enough to be very useful.

What and how one should measure can be determined only by breaking down the general purpose into specifics. A specific purpose of personnel performance measurement is simply a question about that performance which the test planner wishes to answer. In developing specific test purposes, the planner should list every question he wishes the personnel performance data to answer. Each such question becomes a test purpose. Consequently, personnel performance measurement usually has more than one specific purpose.

If there is a possibility that personnel performance will affect other system elements (the equipment configuration, its procedures, environment, or logistics), then a question must be answered. For example, if it is possible that heat/humidity within a tank may degrade tank personnel performance, then the planner wants to accept or reject this hypothesis. Therefore, he must arrange to measure personnel performance in relation to heat and humidity. A specific purpose of the test plan might therefore be phrased as follows:

Example: To determine the capability of tank personnel to operate tank controls while "buttoned up" under high heat/humidity conditions.

This specific purpose tells the evaluator that he must arrange to collect data on operation of tank controls under high heat/humidity conditions. This means that he may have to take the following action:

1. Install a thermometer inside the tank to determine internal temperature/humidity.

2. Measure operator performance as a function of temperature/humidity conditions.

3. Compare that performance with heat/humidity standards specified in MIL-STD 1472B (DoD, 1974).

None of this would have been implied by the general purpose. It is therefore important to list all the specific purposes for which personnel performance measurements will be made.

Section 2.0--DESCRIPTION OF SYSTEM BEING EVALUATED

Describing the system to be evaluated may appear unnecessary to those who are familiar with it. However, in most systems of any size, only some of the system equipment and only some of the tasks involved in operating that equipment will be of interest for personnel performance measurement. These should be identified and described.

If those system operations for which personnel performance data will be collected are not described, investigators may not have a clear idea of how to accomplish the measurement. The larger the system, the more necessary this section is.

It is not necessary to provide highly detailed equipment/task descriptions. If Technical Manuals (TMs) are available at the time of testing, they will provide this detail. If TMs are not available, other system design documents describing relevant equipment should be listed. In any event, it is necessary merely to list the equipment and tasks.

List all equipments for which personnel performance data are to be collected, even if this is only one equipment. For these data refer to TMs and operating procedures. Later on, in describing the criteria and measures the test planner will use (section 4.0), he will consider the characteristics of the equipment listed, because this will dictate his data collection measures.

Example: If the equipment is a radar set, the evaluator may wish to measure number of targets detected or gain level used on the CRT; if the equipment is an artillery piece, he may wish to measure number of rounds fired, speed of loading, etc.

Personnel performance data are collected as part of operational exercises or equipment tests. Ordinarily one would expect to collect such data whenever an operational exercise is conducted but, if this is not possible, those exercises during which the evaluator intends to collect data should be listed. This applies even more so to equipment tests that are being planned, because some equipment tests may not provide a proper environment for personnel performance. A test in which only a vehicle's engines are run up would hardly provide an adequate opportunity to collect personnel data. Because, in almost all cases, personnel performance testing will be "piggybacking" data collection on other tests, the planner should determine which ones he will use and list these. For these data, he should refer to the master test schedule. Whenever the evaluator finds an operation of interest for personnel performance measurement, he should require data collectors to observe the operation and collect relevant data.

Example: If the overall system test plan contains a number of tests which involve significant amounts of personnel interaction, these should be pinpointed as ones for which personnel performance data should be collected.

Not every task performed by operators in a particular test may be of concern to the evaluator. All operations involving operator/maintainer actions should ideally be the subject of personnel performance data collection. If the number of data collectors available is insufficient, however, it may be impossible to record data from all tasks. Under these circumstances, tasks that have the following characteristics should be singled out and identified so that available data collectors can be assigned to them:

1. Tasks that, in the opinion of system engineers or experienced operational personnel, are especially important for satisfactory completion of the system mission.

2. Tasks in which human error or other personnel inadequacies could have a significantly negative effect on the system output.

TMs or other procedural documents describing the tasks for which personnel performance data will be collected should be identified. Hopefully, these documents will contain the personnel performance standards to be used in development of criteria and measures (for use in section 4.0). At the very least, data collectors should be aware of these documents as reference sources and should familiarize themselves with the documents in preparation for data collection.

Section 3.0--SPECIAL COMPARISONS

This section describes Special Comparisons to be considered in performing a personnel performance evaluation.

The basic idea of personnel performance measurement in an operational system test is to determine whether operators can perform their tasks in accordance with system requirements. There may, however, be operational

conditions inherent in the system mission that affect how personnel will perform. The planner may therefore wish to perform special analyses to determine how these conditions affect personnel performance.

Note that these conditions are inherent in system operations and therefore do not require the setting up of special tests to collect required data. For information on such conditions, refer to TMs or other documents describing system operations.

Special comparisons usually imply a comparison of performance under alternative conditions.

> Example: Differences in operator performance between daytime and nighttime operations; between different work shifts; between different EM ranks (when these perform the same tasks).

After such conditions are identified, all one need do is to assign data collectors to observe performance under these conditions and to record appropriate data. If data are to be collected under all mission conditions, comparison data will be collected automatically. However, in writing the Data Analysis section of the Test Plan (section 7.0), the planner should note that a statistical comparison will be made between the conditions of interest. Failure to make such a comparison will reduce the value of the personnel performance measurement. If such comparison conditions do not exist, this section of the test plan should be noted as Not Applicable (N/A). It is assumed that the observations will be made unobtrusively and will therefore not bias the comparison.

The usefulness of such a comparison is that, if operator performance is in fact significantly less effective under one or more of the conditions being compared, analysis of the reasons for this may suggest ways of improving operator performance under those conditions.

Section 4.0--CRITERIA AND MEASURES

Criteria are standards to be met. Two types must be considered: those that describe the performance that personnel are expected to achieve; and those that reflect the adequacy of data collection. The first is more important than the second.

Personnel performance criteria indicate what operators must do in order to satisfy overall system requirements. This information can sometimes be secured from a statement of the system mission or, at a more detailed task level, from operating and maintenance procedures described in TMs or similar documents.

> Example: Operator is expected to record and decode 16 messages per hour; to detect all targets at a distance of 1000 meters.

Criteria are important because they serve as performance standards against which actual operator performance can be compared. This comparison

enables the evaluator to say whether or not the performance of system personnel is satisfactory. If quantitative criteria are lacking, it is impossible to state unequivocally that personnel performance meets system requirements. (The previous sentence does not ignore the usefulness of qualitative data in supplementing the quantitative data. As is explained in Section Three of this report, qualitative data are necessary to help explain the significance of conclusions based on quantitative data; and occasionally no quantitative data can be secured. However, qualitative data alone are insufficient for evaluating personnel.)

Unfortunately, many statements of system requirements do not include quantitative requirements for personnel. Qualitative criteria (e.g., "riflemen will perform reconnaissance missions as quickly as possible") are almost useless as standards, because one cannot compare quantitative data (e.g., mean mission duration was 23.5 minutes) with qualitative criteria. If quantitative criteria are not explicitly specified, it will be necessary for the planner to derive them by using the judgment of experienced military personnel, engineers, or others who are familiar with system requirements (see Development of Quantitative Performance Criteria, p. 2-28. Whatever the source, personnel performance criteria must be described. If there are many detailed performance standards, these can be listed in an appendix to the Personnel Performance Test Plan. At the very least, TMs or other procedures in which the personnel standard can be found must be listed.

Criteria of data collection describe how much data should be gathered. One can collect too little or too much data. The first situation is more serious than the second. With too little data, it is impossible to come to any valid conclusions about performance. Gathering too much data is inefficient, but at least one has the minimum data on which valid conclusions can be based. This is not meant to imply that the more data one has, the more valid one's conclusions.

In the context of the Operational System Test, it is far more likely that too little data will be collected. There are many reasons why this is likely. The time allotted for testing is often too short to permit operators to repeat tasks more than once; equipment may break down and further curtail measurements; emergencies may arise that divert test time to other activities. Under those circumstances, the test director must take what he can get.

However, one principle can be asserted. At least two data items (one repetition of the original event) are required for every operator task for which data are being collected. The size of the data sample needed can be determined by statistical means (see Section Seven of this report). But, in most cases, data collection opportunities will not satisfy that requirement.

The measures (i.e., the types of data) to be recorded should also be specified in the test plan. This will permit the test director to cross-check these measures against the tasks (section 2.3) that require these measures. Failure to list specific measures will often result in failure to record required data.

Measures are derived from two sources: specific test purposes and performance criteria. Test purposes and performance criteria require certain data; and these data call for specific measures. Some criteria and the measures they require are, however, only implied by the system mission. The test director may find it necessary to analyze the logic of system operations in order to extract these criteria and measures. It is also possible to develop measures in the absence of criteria, but such measures may not bear any meaningful relationship to system operations.

Example: During a test, it is possible to record all internal and external communications within a system. In the absence of criteria that specify that personnel should respond to so many incoming calls and make so many outgoing calls within a specified time period, all one can say after analyzing these communications is that personnel did in fact respond to and make so many calls.

Almost all personnel performance measures fall into three general categories: (1) accuracy (or the reverse, errors), (2) duration of responses, and (3) reaction time to initiating events. However, if these are to be useful, they must be specified in terms of the task to be performed and the equipment being operated.

Example: Accuracy (distance from the target in firing the _____ semiautomatic rifle); duration (time taken in loading an artillery piece).

It is also desirable to indicate when the measure will be recorded, if this information is not implicit in the description of the measure.

Measures may vary in their detail, depending on the task level they describe. In firing a rifle, for example, one could record the action of the finger in squeezing the trigger; this would be the most detailed level of measurement. Or one could record the rifleman's error in attempting to hit the target; this would be a higher-order and therefore a grosser measurement level. If a team is involved in the performance one is measuring, the team measure (e.g., whether or not the squad accomplished its objective) would be at still a higher level than measurement of the individual rifleman's performance. The question the evaluator wishes to answer should determine the level at which he will measure. He may wish to measure at several levels simultaneously, but he should measure only when there is a meaningful evaluation question to be answered. Obviously, if he wishes to evaluate squad performance, he would not measure the pressure with which the individual rifleman squeezes the trigger.

The level at which one measures may determine how the data are collected. For example, instrumentation would probably be required if one wished to determine squeeze pressure on the trigger, but not necessarily if one measured error in hitting a target.

Level of measurement is determined by the immediacy of its relationship to the system output. If, for example, the system output one is concerned with is the amount of fire one is placing on a target, then highly

detailed measures like pressure on the trigger would be irrelevant, since squeeze pressure is only indirectly related to amount of fire; rounds expended would be more directly related.

Section 5.0--DATA COLLECTION METHODS

This section deals with who will collect data (e.g., Marines selected from an infantry Battalion; or contractor personnel) and how they will collect it.

Note as a matter of record the number of personnel who are serving as data collectors (section 5.1.1). Where more than one data collector is required, how they are to be scheduled becomes important.

Indicate the task which the data collectors will perform (section 5.1.2). These tasks are not the same as system operating tasks for which personnel performance data are to be collected. Rather, they are the activities involved in gathering information about the task performance being measured.

> Example: At the conclusion of each tank driving cycle, data collectors will administer a questionnaire concerning ride quality.

The level at which these data collection tasks are described need be general only. In the event that there are many data collectors or they have many data collection tasks, section 5.1.2 can be an appendix to the test plan. If several data collectors perform the same data collection task, a single description will suffice.

The reason for describing data collection tasks is to ensure that all personnel know what their responsibilities are. Failure to inform data collectors of their responsibilities will result in lost or erroneous data.

If data collectors must receive special training to enable them to perform their duties effectively, that training should be described (section 5.1.3). In general, all data collectors should receive training in methods of recording data, even if they are specialists in the system being tested. If data collectors are not familiar with the system under test, they will require special training to give them this familiarity, since a data collector cannot function effectively without that information. Data collection training should be oriented toward enabling collectors to recognize the events they have to record and how these should be recorded. At the very least, several data collection "dry runs" with recording forms should be held to habituate data collectors to their tasks.

All data recording forms should be noted in section 5.2. The actual forms themselves should be appended to the test plan. The reason for listing the forms is to remind the test director that these forms must be developed (if they are not already available). Appending the forms to the test plan permits the test director or anyone else to examine them to determine if they are satisfactory for their purposes.

Most data recording tools are of the following types: time and events recording sheets, questionnaires, rating forms, and checklists. Instrumentation used to record data is described in a separate section (5.4).

Unless the data collection procedure is very simple, it should be described in some detail (section 5.3). This description is to let data collectors and all others involved in the evaluation know what is required of them. It should include the following (as relevant) and any other significant aspects of the work:

1. The hours data collectors will work, or the sequence of operations (their beginning/completion) that will determine the data collection period.

2. How data collectors should process their data. For example, do they pass the data on to the evaluation personnel immediately or hold on to it? Do they do any analysis of the data during the performance being evaluated?

3. What data collectors should do if an emergency occurs (e.g., if an exercise is suddenly stopped before it is completed), or if something not covered by operating procedures occurs.

4. The level of detail to which they record data (principally relevant when reporting qualitative observations).

5. The extent to which the data collector is permitted to interact with the personnel whose performance is being evaluated (e.g., the distance they must remain away from participants in the operation).

6. Any equipment data collectors will be required to operate.

Although there are occasions when elaborate instrumentation will be required for personnel performance measurement, in general, instrumentation demands should be minimized, particularly if testing is being conducted in a field environment. It is difficult to operate elaborate instrumentation effectively in that environment, especially when the equipment requires highly controlled conditions.

Data collection instrumentation that is sometimes used in field operations include small magnetic tape recorders for recording communications and hand-held video tape recorders or motion picture video tape cameras for recording events visually. To assess the environmental conditions under which performance occurs, light meters and accelerometers (for vibration effects) may be employed. Sound level meters may be used to record noise levels. Such instrumentation should be used only if data on environmental conditions are necessary and have not been secured in earlier developmental tests. If evaluations are being held in fixed locations, either on the ground or aboard ship (i.e., where greater control can can be exercised over environmental conditions), more elaborate automatic data recording apparatus may be used.

Section 6.0--SUBJECTS

This section describes the subjects whose performance is being evaluated; that is, those personnel who are selected to operate and maintain the system during its testing (also referred to as test personnel).

If the operational system test is to be valid, it must be performed with personnel who are representative of those who will eventually operate and maintain the system. If, for example, subjects are much more or less trained or experienced than eventual users of the system, evaluation results will fail to describe correctly the performance of these eventual users. If the evaluation is performed under military auspices, selection of the appropriate subjects should pose less difficulty than it would if it is performed with contractor subjects. Even under military conditions, however, the evaluator must exercise care in the selection of subject personnel. At the very least, he must examine any special requirements that user personnel will have, and match his subject characteristics to those requirements.

Test personnel requirements include the following:

1. Physical--e.g., vision (20/20), hearing, height, weight, and strength.

2. Aptitude--general intelligence; special aptitudes (e.g., mechanical).

3. Training--e.g., graduation from a required training course.

4. Experience--number of years in a given military speciality.

Attitudinal and motivational factors are important also but cannot be precisely specified. If military personnel are used as subjects, they are, of course, subject to orders; but, even in this case, it is important to explain to personnel why they are being asked to participate in the OST and the importance of their performance.

Most frequently, subject requirements will involve training and experience. For example, it is obvious that, to evaluate a prototype tank, the subjects must be qualified tank drivers and, if driving the new tank requires new skills, they must have received factory training in tank driving. Although the personnel requirements in the example are relatively obvious, others may not be. In such cases, the evaluator may have to determine them from analyzing the personnel requirements information in documments describing the system.

Section 6.1 notes the number of subjects or the organizational units in which they function (e.g., the squad, the platoon). This item is for information only.

Section 6.2 describes required subject characteristics. These are characteristics that, in the evaluator's judgment, will significantly affect the accuracy of the data if not possessed by personnel acting as subjects.

Section 7.0--DATA ANALYSIS

It is essential that the test plan include a procedure for statistical analysis of the data. If the test director waits until after data are collected before developing his statistical plans, the chances are excellent that too much or too little data will be collected and--much more serious--much of his data will be unusable because it will not fit statistical requirements.

All data analysis in system testing involves one or both of two types of comparison. The first compares personnel performance in the system under evaluation with a system requirement; this comparison verifies that personnel can perform required tasks (general purpose, section 1.1). The second comparison is between any special conditions that are tested (e.g., daytime vs. nighttime performance) (see section 3.0). In both cases, the evaluator needs a statistical plan that describes the kind of statistics he will use in relation to which data measures in order to answer which questions.

Because planning an appropriate statistical data analysis is a highly specialized technique, it is advisable for the test planner to allow a statistical expert to prepare this section of the test plan (unless, of course, he is also a statistician). At least the planner should permit an expert to review the statistical design of the study and the entire test plan before it is finalized.

Section 8.0--TESTING SCHEDULE

This section describes the personnel performance data collection schedule. If the system test is very complex, and only some of the system operations will be used as occasions for the gathering of personnel performance data, then a daily, weekly, or monthly schedule of data collection activities should be appended to the test plan.

Example: In one flight navigation testing program for helicopter pilots, pilots of varying levels of experience were to be tested over a year's time. Since pilot navigation performance was hypothesized to depend in part on the appearance of the terrain, it was necessary to systematically arrange the subject schedule so that pilots with different experience levels could be tested during both summer (heavy foliage) and winter (bare trees, snow) conditions.

If the personnel performance data are to be collected during all test events, a detailed subject schedule is unnecessary because the overall test schedule will determine where data will be collected.

Example of Personnel Performance Test Plan--Personnel Performance Measurement in the XM-47 Amphibious Troop Transport and Tank Destroyer

The example of a Personnel Performance Test Plan described in the following paragraphs has been written to illustrate the preceding instructions and to serve as a model for such a test plan. It is based on an

imaginary man-machine system, the XM-47 amphibious tank/transport, which has some of the characteristics of the Marine Corp's LVTP-7 but which does not pretend to any realism in terms of physical engineering details or operational capabilities. The reader should consider this example as representing a type of system the Marine Corps might have in the future (e.g., 1985).

The XM-47 is an amphibious troop transport that also performs land reconnaissance and functions as a tank-destroyer. It has a crew of three: a driver, a missile loader and mechanic, and a gunner. It can carry 10 fully loaded troops or an equivalent amount of supplies from a "beyond the horizon" mother ship to shore, disembark these troops or supplies, and then perform ground reconnaissance and/or antitank duties. It is armed with the following weapons: (1) Typhoon heat-seeking surface to air (SAM) missile for protection against aircraft (range up to 10,000 feet), (2) the Hurricane antitank guided missile (range up to 2 miles), and (3) two 50-calibre machine guns. It has a top speed of 20 knots in water, 50 mph on well-surfaced roads, and 30 mph over sand and fields.

Testing of this system will be performed as part of an amphibious exercise involving adversary units, including troops, artillery, and aircraft.

The material to be included in each of the various sections of the test plan is described below. The author's comments are presented in italics and indented form. These comments would not ordinarily be found as part of the test plan.

Section 1.0--PURPOSE

1.1 General. The purpose of this test is to verify that system personnel can perform all required tasks to accomplish mission objectives.

1.2 Specific.

In any highly advanced system like the XM-47, many questions beyond the general purpose need to be answered. For example, although considerable analysis went into the specification of a crew of three for this vehicle, it is possible that because of workload considerations more personnel will be needed; hence, only realistic simulated-combat exercises will verify that a crew of three is adequate. Similarly, although noise vibration data were gathered as part of developmental testing, these data did not include measurement of personnel effects resulting from the tank environment; hence, the effect of environmental conditions on performance effectiveness remains to be tested. The adequacy of operating and maintenance procedures can be tested adequately only under operational conditions; hence, this question is also pertinent. To repeat what was said previously: every question which is desirable to answer during this test becomes in effect a specific purpose of the test. (One assumes, of course, that such questions are not frivolous.)

The following are the specific purposes for which the personnel performance evaluation will be conducted:

1. To determine whether crew size is adequate to perform all required tasks without overloading personnel excessively.

The qualification "without overloading personnel excessively" is necessary because in many cases personnel can adapt to stressful or undesirable situations but at the cost of increased likelihood of error. Although such stressful tasks may be accomplished, one may not wish to accept the added risk of error.

At the same time, note that there is a certain ambiguity about the qualification "excessively." It will be necessary for the test planner to determine and specify a criterion of acceptable performance (e.g., number of errors) under overload conditions.

2. To determine whether environmental conditions (heat, noise, lighting, vibration) within the XM-47 will permit operators to perform all required tasks without excessive overload.

3. To determine whether the environmental conditions noted in (2) above will permit troops when transported to shore to perform necessary combat functions without loss of efficiency.

4. To determine that XM-47 missile weapons can be fired adequately.

This implies all the personnel functions involved in successful operation of XM-47 missiles: detection, identification, and hit probability. Again, it is assumed that quantitative criteria for these functions exist or can be determined.

5. To determine that operating procedures can be performed correctly and adequately.

Criteria for these should be specified in section 4.1.

6. To determine the frequency and types of operator errors made and their causes.

7. To determine that the driving characteristics and "rideability" of the XM-47 are adequate.

8. To determine the effects of any fatigue occurring during a routine mission.

9. To determine the adequacy of seat design in the XM-47.

The above list of purposes is not necessarily complete; it is only illustrative. Moreover, these specific purposes will require particular measures and measurement procedures.

The above purposes can, if desired, be broken down even further. For example, many elements make up adequate task performance, such as reaction time to enemy attack or a maximum duration for performing a specific task. Should these elements be pulled out as specific purposes of the test? Not necessarily, as long as the measures developed to satisfy these purposes are detailed enough to provide desired data.

Section 2.0--DESCRIPTION OF SYSTEM UNDER EVALUATION

A general description of the system (e.g., the previous paragraphs describing the XM-47) need not be provided here, since it is reasonable to assume that all concerned are familiar with the general characteristics of the system. However, the specific equipments and tasks with which the personnel performance evaluation is concerned would be listed.

2.1 Equipments Operated (Maintained) by Personnel. Personnel performance data will be collected with regard to the following equipments:

a. Controls and displays for driving the XM-47 (see TM listing below for equipment details).

b. Optical displays for locating ground targets and for control during firing of the Hurricane antitank missile.

c. Controls for loading and firing the Hurricane antitank missile.

d. Radar display for detection of aircraft targets.

e. Controls for firing Typhoon surface to air missile.

f. Radio set 4FU.

g. .50-calibre machine gun.

h. Operator seats and restraint harnesses.

Evaluations will be performed with regard to the following questions: Are the seats comfortable for prolonged missions; and do they provide sufficient arm/leg room to operate controls?

i. Internal XM-47 environment.

The definition of equipment operated by personnel extends to the working environment. It requires consideration of the work space available to operators, controls for adjusting lighting and temperature, etc.

Implications of the items in this section for the evaluation include use of human engineering checklists (see section four) to evaluate arrangement of controls and displays; and personnel responses (in the form of questionnaires and rating scales) concerning ease/difficulty of operating controls and the driving qualities of the XM-47.

2.2 Equipment Tests During Which Personnel Performance Data Will Be Collected. Personnel performance data will be collected during amphibious/ground attack exercises (Blue Force) conducted the week of March 2nd, 1985 at Ocean Peninsula. Data will be gathered on all XM-47 operating cycles. Two prototype vehicles will be utilized, alternating on missions. XM-47 will be subject to control of CO, Blue Force. Two XM-47 crews comprised of three men each will be used in alternation. Start of an XM-47 cycle will be positioning of the vehicle in the well of an LSD. Conclusion of the cycle will be the recall signal as received over XM-47 radio. Approximate duration of the XM-47 mission: 3 hours. Total number of XM-47 mission cycles: approximately eight (two missions per 24-hour interval). On half of these missions, the XM-47 will be loaded with 10 fully equipped troops; on the other half, with supplies. Loading and off-loading of the troops/supplies will be included in mission functions being evaluated. Upon command, XM-47 will drive from the LSD to shore, disembark the troops or supplies, and will conduct reconnaissance/antitank operations against an enemy force of infantry, artillery, tanks, and aircraft. Half the XM-47 operations will be conducted at daytime, the other half at night.

This description has certain implications for the design of the Human Factors evaluation. The fact that the evaluation will take place during simulated combat operations means that certain unprogrammed events (such as enemy air attack) will occur which the data collector must record. Since two vehicles and two different crews will be used in alternation, a statistical comparison of performance between the two vehicles and crews must be conducted. If, for example, statistically significant differences between the performance of the two crews are found (which is unlikely but might occur if the two crews were not given equivalent training), then data cannot be combined and an analysis must be performed to determine why the differences occurred. Since the missions will be divided between troop carrying and supply transport, a comparison of performance between these two types of missions is required, although on the surface it would not seem as if

this factor alone would be important. The fact the XM-47 operations will be divided between daytime and nighttime also requires a statistical analysis to determine if performance differs significantly between these two conditions, and if so, why. Since different geographical terrain conditions are involved (sea, shore, sand, hard surface road, and field), comparisons of performance under these conditions must be performed. Since approximately eight missions will be conducted, mission-to-mission performance should be analyzed. Questions addressed to test personnel will have to be developed in the light of each of the conditions entering into the evaluation.

2.3 Tasks For Which Personnel Performance Data Will Be Collected. Data will be collected concerning performance of the following tasks:

a. Premission checkout.
b. Driving the XM-47.
c. Loading/unloading troops/supplies.
d. Communications over radio and intercom.
e. Operation of the .50-calibre machine gun.
f. Detection of ground and air targets.
g. Firing at ground and air targets.
h. Any required corrective maintenance during the mission.
i. Postmission checkout.

This list ties the evaluation to specific personnel operations. For example, pre- and postmission checkouts are mentioned here for the first time. If such a list were not developed, it is possible that important tasks would be overlooked by the data collector. This list can also be compared against the data recorded and thus serves as a check on the completeness of those data.

2.4 Applicable TMs and Other Documents. The following list of TMs applies to the equipments and tasks being evaluated:

a. TM _____: Checkout Procedures for the XM-47.
b. TM _____: Operator's Manual for the XM-47.
c. TM _____: Operation of the Typhoon surface-to-air missile.
d. TM _____: Operation of the Hurricane Antitank Missile.
e. TM _____: Operation of the 4FU radio set.
f. TM _____: Operation of the .50-calibre Machine Gun.

Other applicable documents include:

a. Rpt. _____: Mission Requirements of the XM-47 Transporter/Antitank Vehicle.
b. Rpt. _____: Function Analysis of XM-47 System Operations.

This list tells data collectors which documents they should refer to in order to secure information about these individual weapons and the features of equipment operation tasks which they should record. It can serve also as a source of information concerning criteria and measures (see section 4.0). Data analysts will probably find these documents useful during the data analysis and writing of the final test report. All data collectors should be issued with a copy of applicable TMs.

Section 3.0--SPECIAL COMPARISONS

The following major comparisons will be made:

a. Performance under daytime vs. nighttime conditions.
b. Between vehicles.
c. Between crew 1 and crew 2.
d. Between troop carrying and supply missions.
e. Between successive operations (mission 1 vs. 2 vs. 3, etc.).
f. Driving qualities on water vs. sand vs. surfaced road vs. field.

This list clarifies the specific analytic comparisons to be made. Previously they had merely been implied by section 2.2. Note that these conditions are inherent in the mission and should not require special tests for their comparison.

Section 4.0--CRITERIA AND MEASURES

4.1 Criteria.

The following section presents only a sample of the criteria which should be applicable to XM-47 operations. For example, criteria to represent acceptable performance under overload have not been indicated, although they should be, since one of the purposes of the evaluation is to determine that the crew can perform effectively under overload conditions.

The following criteria will be used to verify that system personnel can perform required tasks.

a. Maintenance of an average driving speed in water of 12 knots; in sand, 22 knots; on road surfaces, 35 knots; over fields, 18 knots.

b. Maximum unloading speed: personnel, 3 minutes; supplies, 5 minutes.

c. Probability of hitting a stationary ground target: .95.

d. Probability of hitting a moving ground target: .85.

e. Completion of all procedural tasks within accuracy and time limits specified in applicable TMs.

f. Not more than one discernable instance of incapacitating motion sickness per mission among troops being transported.

g. Ride quality and driving characteristics adequate on all road surfaces.

These last criteria will require quantification before they can be meaningfully measured. It is apparent that much more data must be collected in the personnel performance evaluation than is represented by the preceding criteria, although each datum ideally should be referrable to a specific criterion.

4.2 Measures.

Measures are broken down according to the following categories: performance data, environment, equipment-human compatability, and subjective data. The list of measures is incomplete and illustrative only.

The following data will be recorded:

a. Performance data

(1) Start/stop time for each mission.

(2) Start/stop time for loading/unloading troops/supplies.

(3) Instances of incapacitating motion sickness in troops being transported.

(4) All internal and external communications.

These will require more detailed analysis at a later time; we do not presently have specific criteria for communications analysis.

(5) Start/stop time for premission equipment checkout.

(6) Start/stop time for postmission equipment checkout.

(7) Start/stop time for all instances in which XM-47 attacks a target or is itself attacked.

(8) Number of rounds fired and hits against surface targets.

(9) Number of rounds fired and hits against aircraft targets.

(10) Type and number of instances of error in performing procedural tasks.

Note that without instrumentation it will be difficult for an observer to note all errors; hence only major errors with significant effects on personnel will be recognized and recorded.

(11) Average driving speed on various types of surfaces.

(12) Start/stop time for all contacts with targets.

Where these are very frequent, it may be difficult to get highly precise data.

(13) Start/stop time (downtime) resulting from all equipment malfunctions.

(14) Actions taken to restore equipment to functioning status.

Assuming that observational data recording methods only are available, this information will not be very detailed (e.g., who worked on the equipment, major actions taken).

(15) Distance error in arriving at beach head location.

b. Environment

Noise, lighting, vibration, temperature measures extracted from MIL-STD 1472B (Department of Defense, 1974).

c. Equipment-Human Compatability

(1) Adequacy of equipment layout for operator use, including major human engineering discrepancies.

(2) Aspects of equipment layout and operating procedures which should be modified.

These are more properly recommendations for corrective action. However, since the observations and deductions on which these recommendations are based must be secured during the actual test, they are listed here.

(3) Discrepancy report: All discrepancies from optimal functioning (including equipment failures) will be recorded at the close of the mission.

d. Subjective data

(1) Operator rating of adequacy of environmental conditions and suggestions for modification.

(2) Troop evaluation of ride comfort.

(3) Driver evaluation of driving and ride characteristics of XM-47.

(4) Operator evaluation of aspects which require additional emphasis in training.

(5) Tank personnel evaluation of adequacy with which the total mission was accomplished.

Subjective data collection techniques will be described in Section Three of this manual.

Section 5.0--DATA COLLECTION METHODS

5.1 Data Collectors

Data collectors are emphasized in the Test Plan because, in an operational system test conducted in the field, automatic methods of data collection may not be feasible.

5.1.1 Number. Two (2) data collectors will accompany each XM-47 test mission. The reason for having two data collectors is to check the reliability of the observations made. Should it be determined, on the basis of initial tests, that one data collector will provide reliable data, the number of data collectors will be reduced.

5.1.2 Data Collection Tasks. Data collectors will record the start/stop time of all discrete functions that are described in section 4.2. Where instrumentation is required for recording personnel performance data (e.g., tape recorders for communications), they will activate and monitor the functioning of the instruments. During the test, they will record all significant events by maintaining a running diary (see appended forms). All data recording will be on the basis of noninterference with on-going mission tasks. At the conclusion of the test mission they will administer, collect, and process all paper-and-pencil data forms.

5.1.3 Data Collection Training. It is assumed that the data collectors will have been given factory training on the XM-47. Specific data collection training will take the form of dry runs to familiarize personnel with data collection procedures. It is estimated that at least two (2) dry runs, corresponding as much as possible to the actual test operation, should be made. During the dry run, data will be collected as it would be in the test. Following each dry run, the data collected will be examined and a critique made by the Personnel Performance Data Evaluator.

The nature of the training to be provided data collectors depends ultimately on the functions assigned them. This, in turn, depends at least partially on the system being evaluated. Rarely will a formal training course be required, but a dry run such as indicated above will always be necessary unless the data collection task is extremely simple.

5.2 Data Collection Forms. The following records are appended to this test plan:

a. Time and event records.
b. Instrumentation data records.
c. Subjective data records (questionnaires, rating scales).

5.3 Data Collection Procedures. These have already been described in section 5.1.2 (Data Collection Tasks).

More detailed data collection procedures will be written only when necessary.

5.4 Instrumentation. The following equipment will be installed aboard the XM-47:

a. Sound level recorders.

b. Accelerometers for measurement of motion and vibration.

c. Magnetic tape recorder for continuous recording of crew communications.

d. Equipment for automatic recording of driving speeds.

e. Laser target recorder attached to missile firing equipment for automatic recording of firing and hits.

The Appendix to this test plan presents drawings of equipment locations and attachment details.

Data collection personnel will be responsible for monitoring the operation of the equipment, periodic recording of data values, and performing first-level maintenance should a failure occur.

Failure of instrumentation during the test will not be cause for cancellation of a test mission.

Section 6.0--SUBJECTS

6.1 Number. Subjects are of two types: (a) the XM-47 crew of 3, and (b) troops carried (10).

6.2 Required Characteristics. Selection of the XM-47 crew is the responsibility of the Test Director. Crew members must meet basic qualifications of the following MOS: _____. In addition, they must have satisfactorily completed the contractor's factory training course in operation of the XM-47. They must also be proficiency-certified in firing the Hurricane antitank missile and the Typhoon SAM.

The troops carried in troop-carrying missions need no special qualifications.

Section 7.0--DATA ANALYSIS

The following analyses will be performed:

a. Determination of mission duration.

b. Determination of maximum time for loading and unloading troops and supplies.

c. Determination of pre- and postmission checkout duration.

d. Proportion of ground and air targets detected and correctly identified and compared with criterion.

e. Proportion of ground and air targets hit and compared with criterion.

f. Comparison of average driving speeds per type of surface.

g. Number of equipment malfunctions per equipment and downtime per equipment.

h. Average error in arriving at beach head.

i. Number of instances of incapacitating motion sickness in troops being transported.

j. Determination of mean and maximum temperature during mission.

k. Determination of mean and maximum vibration during mission and comparison with standards.

l. Determination of mean and maximum noise level during mission and comparison with standards.

m. Instances of major error in operation of equipment.

n. Evaluation of adequacy of equipment layout.

o. Evaluation of adequacy of environment conditions and ride comfort.

p. Evaluation of adequacy of driving characteristics.

q. Subject recommendations for modification of equipment design, procedures, and training.

r. Comparison of tank crew performance under day and nighttime conditions.

s. Comparison of tank crew performance in vehicles (1) and (2).

t. Comparison of tank crew performance under troop carrying and supply conditions.

u. Comparison of tank crew performance over succeeding missions (mission (1) vs. (2) vs. (3), etc.).

Note that in these last comparisons the measures to be employed will make use of most if not all the measures listed in section 4.2.

It will be seen by comparison with section 4.2 that the data analysis consists largely of a compilation of the measures in that section. No very sophisticated experimental design is required because this operational test is not concerned with the examination of variables except for the comparison of day vs. nighttime operations, etc. In general, the conditions under which operational system tests are conducted do not permit elaborate experimental designs.

Section 8.0--TESTING SCHEDULE

The personnel performance test schedule is concurrent with the overall test schedule. No special test missions are required for the collection of personnel data. Performance data will be collected during all tests in which personnel operate equipment in an operational manner.

Development of Quantitative Performance Criteria

One of the required tasks in developing the test plan is the specification of personnel performance criteria. Without these, no meaningful evaluation of personnel performance or the system as a whole is possible. Ideally, such criteria should be available in function/task analysis documents and TMs accompanying the prototype system to be evaluated. Unfortunately this is often not the case, and so the resultant evaluation--although providing useful information about the system and how it should be improved--does not accomplish the purpose of the evaluation; i.e., verifying that personnel performance satisfies system objectives. This substantially reduces the value of the operational system test.

Even when quantitative criteria are not available, it is possible to derive them, or at least to develop approximate values of such criteria. The purpose of this section is to describe a systematic procedure for deriving these quantitative values. The procedure described is based in part on what has been termed the Delphi technique.[2] Since performing this procedure will take time and effort, the development of the Personnel Performance Test Plan cannot wait until shortly before the evaluation is to begin.

---

[2]For information on the Delphi technique, see the following: Dalkey, N., and Helmer, F. An experimental application of the Delphi method to the use of experts. Management Science, 1963, 9, 458-467 and Sander, S. I. Delphi: Characteristics and applications (NPRDC Technical Note 76-2). San Diego: Navy Personnel Research and Development Center, October 1975.

At this point, it is desirable to digress and distinguish between overall system criteria and personnel performance criteria--a distinction which was made previously but which bears repetition.

Example: An example of a system output criterion is the following: Accuracy of the artillery piece should be ±50 yards. An example of a personnel criterion is: Accuracy in laying (pointing) the piece should be ±10 mils.

It is important for the evaluator to be able to distinguish between system and personnel performance criteria because he may otherwise fail to measure personnel performance correctly. System criteria describe what is required of the system as a whole, which includes personnel performance; personnel criteria deal only with human functions in the system. In the above example, the measurement of artillery accuracy includes not only the accuracy with which the piece is laid, but physical characteristics of the gun itself, windage conditions, etc. It may appear as if measurements of the system criterion would be sufficient for test purposes, because it includes the effect of personnel performance. However, to fail to measure the human function individually would be to lose much of the value of the system measurement. Suppose, for example, that accuracy of the rifle were actually ±100 yards, which is unacceptable. What is the cause of the inaccuracy?

The major factor causing the unacceptable inaccuracy might be failure of the artillery battery to lay the gun correctly. If one did not measure the human function separately and then relate the human error to the system error, one would never know what needed to be done to bring the error within acceptable limits. One of the major goals of personnel performance testing is to relate personnel error to system performance; one cannot do this except by measuring both factors.

The reason for the lengthy discussion of system vs. human criteria is that system criteria are much more likely to be expressed quantitatively in system documents than are human criteria. As a result the test planner may be tempted to concentrate on the system criteria and to ignore the human ones.

One of the first steps in development of the test plan is to examine the available system data and determine how much criterion information is available. In the case of the system as a whole, the planner should list all the functions to be performed by the system, and then determine for which functions quantitative data are available. He should make a similar list of personnel functions and tasks and the quantitative standards available for these.

Where does this list come from? Theoretically, function/task analysis documents are produced during system development and should accompany the prototype system. Review of these documents should give the test planner at least the major personnel functions and tasks. The documents may not provide a complete listing, however, and the planner will then have to supplement the documents.

The general strategy for deriving a list of tasks together with their quantitative criteria is to make use of a group of subject matter experts. A subject matter expert is someone highly experienced in utilizing a system of the type being evaluated. For example, an officer who has several years' experience in tank operations might well be considered a subject matter expert for a new tank.

There is nothing unusual in the use of subject matter experts to assist in the development of a new system. The Delphi-Like procedure described below is simply a more systematic way of securing certain judgments than might otherwise be secured.

Step 1: The test planner will recruit a number of subject matter experts. Because of the variability inherent in subjective judgments, even from highly experienced personnel, it is best to have as large a number of experts as one can find. Practical considerations enforce an upper limit on the number of these experts. It is recommended that no less than three nor more than five experts be recruited.

Step 2: The test planner will present to the group of experts a list of tasks to be performed by operators of the system to be evaluated. Each expert will be asked to review this list individually. (The reason is that the planner wishes each expert's opinion to receive full weight. In a group discussion, higher-ranking experts may dominate lower-ranking ones; those with more forceful personalities may overawe those with less aggressive personalities.) Clearcut instructions should be written so that each expert knows exactly what he is supposed to do. Each expert will be asked to answer the following questions:

a. Is the list of operator tasks complete? If not, add whatever tasks you consider important to accomplishment of the system mission.

b. Is the level of the tasks described meaningful in terms of accomplishing the task?

Refer back to the previous discussion on this topic. Remember that some tasks may be so grossly described that they include (but do not specify) a number of subtasks for which data should be collected. These latter should be specified. On the other hand, certain tasks may be so molecular that their measurement would be pointless in terms of the overall goal of accomplishing the system mission.

If the task is not described correctly, have the expert rewrite the task as he sees fit, with emphasis on the meaningfulness of measuring that task.

Step 3: After each expert individually has reviewed and modified the list of tasks, collect the revised lists. Then the planner will combine all the lists into one, noting which tasks have been eliminated and which have been added.

Step 4: At this point, reconvene the experts--this time as a group. Give each one the combined task list. Have them examine the list and attempt to agree on that list. If agreement cannot be reached on any task, eliminate that task on the assumption that any such task is probably not of significant importance to mission accomplishment.

Step 5: Once a composite list is agreed upon, ask the experts to take that list and now--again individually--rank-order the tasks in terms of importance. Three criteria of importance ought to be kept in mind by the experts when they rank-order the tasks:

a. Importance in terms of accomplishing the system mission.

b. Importance in terms of system safety (e.g., maintaining the integrity of the system in the face of enemy fire).

c. Importance to the safety of operator personnel.

The reason for prioritizing the tasks in terms of importance is that it may not be possible to measure all of them during testing. The number of data collectors may be too few to collect data on every test or for every task in a given test exercise. Certain tasks may be performed concurrently, which means that two or more data collectors will be required. Some of the tasks may be of such relatively slight importance that there is no pressing need to collect data on them. Of course, if the number of tasks is sufficiently small, and they are all equally important (unlikely but possible), then the step of prioritizing tasks can be eliminated.

At the same time, in the interest of reducing the claim upon the time of the subject matter experts, they should be asked to indicate for each task the quantitiative criterion to which the task should be performed.

> Example: Task (in connection with artillery forward observers (FO): Locate enemy position by reference to map/ grid coordinates. Criterion: ±100 yards.

Selection of a criterion value should be based on operational logic. If, for example, the task is to locate an enemy battery for counter-battery fire, the criterion standard should not be more precise than counter-fire accuracy would demand. If artillery cannot fire more accurately than ±100 yards, there is no point in requiring the FO to locate the enemy target to ±50 yards. On the other hand, if artillery fire accuracy is ±50 yards, one should require the FO to locate to the same value (provided that this is possible within the FO's perceptual capabilities).

Step 6: After the subject matter experts have individually prioritized the tasks and supplied quantitative criteria (qualitiative criteria should not be accepted, since these cannot be used for evaluation purposes), the planner will collect the lists and criteria and develop a composite product.

Step 7: Call the subject matter experts together as a group and have them review the composite list which indicates all the priorities and

criterion values specified previously. As in preceding steps, the experts will be asked to thrash out all disagreements and develop a list on which they all agree. Where the experts cannot agree on a task priority or a criterion standard, take the median value of the judgments made. At this point, the work of the subject matter experts is completed, at least for the job of prioritizing tasks and establishing criterion values. However, these experts may be utilized later in reviewing the data analysis.

Step 8: On the basis of the task/criterion listing and using as a selection criterion the number of data collectors he has available, the test planner will select whatever number of tasks he considers as feasible to measure. The criteria specified for those tasks will be noted and made available to data collectors for use in their measurements.

In complex tasks the task may have multiple dimensions. For example, in air traffic control, the following performance criteria may all be important: (a) distance between aircraft landing; (b) number of aircraft in a landing pattern; and (c) number of aircraft landed in a given time period. The test planner must ensure that subject matter experts do not ignore some relevant task dimensions for which a standard must be available.

In evaluating the importance of the tasks, the planner may wish to have subject matter experts indicate the absolute degree of importance the task has, rather than (or in addition to) its rank order. The expert may be asked to rate each task on a scale such as the following:

Example:

1_____2_____3_____4_____5

1. Critical to system functioning.
2. Extremely important to system functioning.
3. Very important to system functioning.
4. Somewhat important; can be delayed, but should be performed.
5. Relatively unimportant; in an emergency may be skipped without affecting system operations.

The planner may also find certain tasks whose standard of performance depends not on the individual operator but, rather, on circumstances outside the operator's control (e.g., other systems or the enemy's action). Such tasks present special problems in setting up a quantitative performance criterion.

In establishing a criterion for such tasks, the experts will have to estimate the maximum number of inputs to the operator that other systems or the enemy are likely to provide. For example, in establishing the number of coded messages which an intelligence operator would be expected to receive, decode, and transmit to G-2, the expert must consider, on the basis of his experience, the highest likely workload which the operator may have and the maximum amount of work he can be expected to perform under these circumstances. This is not a very satisfactory principle, but it is the best one that can be established for this type of uncertain situation.

The criteria thus established are, of course, not "fixed in concrete"; actual test experience may cause them to be modified.

SECTION THREE--THE SELECTION AND DEVELOPMENT OF MEASURES AND MEASUREMENT METHODS

The measures and methods described in this section can be used for both individual operator and team evaluations. Ideally, the most appropriate measures are objective, quantitative, unobtrusive, and easy to collect, require no specialized data collection techniques or instrumentation, and cost little or nothing. Objective measures are preferable, but often cannot be used. Subjective methods are quantifiable and can be used to supplement objective ones. The objective measures described include reaction time, duration, accuracy, frequency of occurrence, and amount achieved. The subjective methods available include the interview, questionnaire, observation, ratings, checklists, and critical incidents. This section describes the information each measure/method can supply, under what circumstances it can and cannot be used, its advantages and disadvantages compared with other measures/methods, and factors to be considered in selecting the measure or method.

Introduction

The personnel performance test planner has a major responsiblilty to select the measures and measurement methods with which data will be collected. Initially, the test planner selects only a general class of measure (e.g., reaction time); thereafter, he must develop a specific measure by applying the general measure to the particular task and system he is evaluating. For example, he may decide that he wishes to measure an equipment operator's reaction time. He must then decide when he should begin timing performance and the specific response that represents the end of the task period being timed. On the other hand, if he wishes to assess accuracy, he must first define what accuracy (or its converse, error) is in the context of the particular task and system.

Much theoretical material has been written about individual measurement methods and ways of utilizing them. However, since this manual is designed to be used in the actual conduct of an operational system test (OST), only the most important elements of this material have been included; matters of theory which cannot be immediately applied to the OST have been largely ignored. The test planner should, if at all possible, have the assistance of a human factors specialist in developing his measurement methods.

Examples

In order to describe these methods concretely, an actual Marine Corps system--the Tactical Air Communications Central (TACC) (AN/TYQ-1)--will be used as a continuing example throughout the remainder of this section. The following is a short description of this system, its equipment, and the functions performed by its operators. (More detailed information can be found in Marine Corps Technical Manual TM-04428A-1005-14/1.)

The TACC system is part of the Marine Tactical Data System, which is a landbased air control facility. The particular subsystem of the TACC used as the example is the Operations Section, which has five subsections: command, air defense, air support, traffic, and display. It requires 19 operators, although only 11 of these are key personnel.

The major equipments operated by these personnel are five situation display consoles, six communications desks, a Weapons Availability Status Display Group (WASDG), eight manual status boards, and a map board. A computer generates alphanumeric and vector (line) video information for display on the situation display consoles and WASDG, and continuously updates these to provide operators with current tactical information. The situation display consoles contain information needed to evaluate the tactical situation such as a reference map, friendly and hostile aircraft symbols, projected track of aircraft (lines), and reference and identification data (alphanumerics). The operator can display elements at different intensities or remove display elements from the CRT screen. He can select different scales for the display and change the center of the display to focus on a particular area of interest.

The WASDG displays alphanumeric information concerning targets and interceptors (engaged or nonengaged), fire units available, and miscellaneous friendly resources. This is for group viewing.

The Operations section also includes a total of 10 intercommunication stations with two-way voice communications between operations and planning group personnel. These communications desks interface with remote field telephones and radio. A teletype permits secure or nonsecure teletype communications with remote facilities; it provides relative low-speed, textual traffic. A loudspeaker-amplifier permits public address of any voice communications signal available at the communications desk.

TACC personnel perform a wide variety of operational and behavioral functions. The operational functions include the following:

1. Monitor, supervise, and coordinate the control of aircraft for air and assault support.

2. Control the launch and allocation of on-call air support aircraft.

3. Coordinate all air traffic in area of responsibility.

4. Monitor equipment status and operational posture of other Marine Corps organizations relative to air support.

The behavioral functions include:

1. Detection of video signals.

2. Operation of control consoles.

3. Analyses of alphanumeric and symbolic data.

4. Verbal communication.

5. Interaction with others.

Although operations are governed by superordinate procedures, they are somewhat unstructured at the working level, flowing from immediate needs.

The configuration of Operations section stations in their inflatable shelter is shown in Figure 3-1.

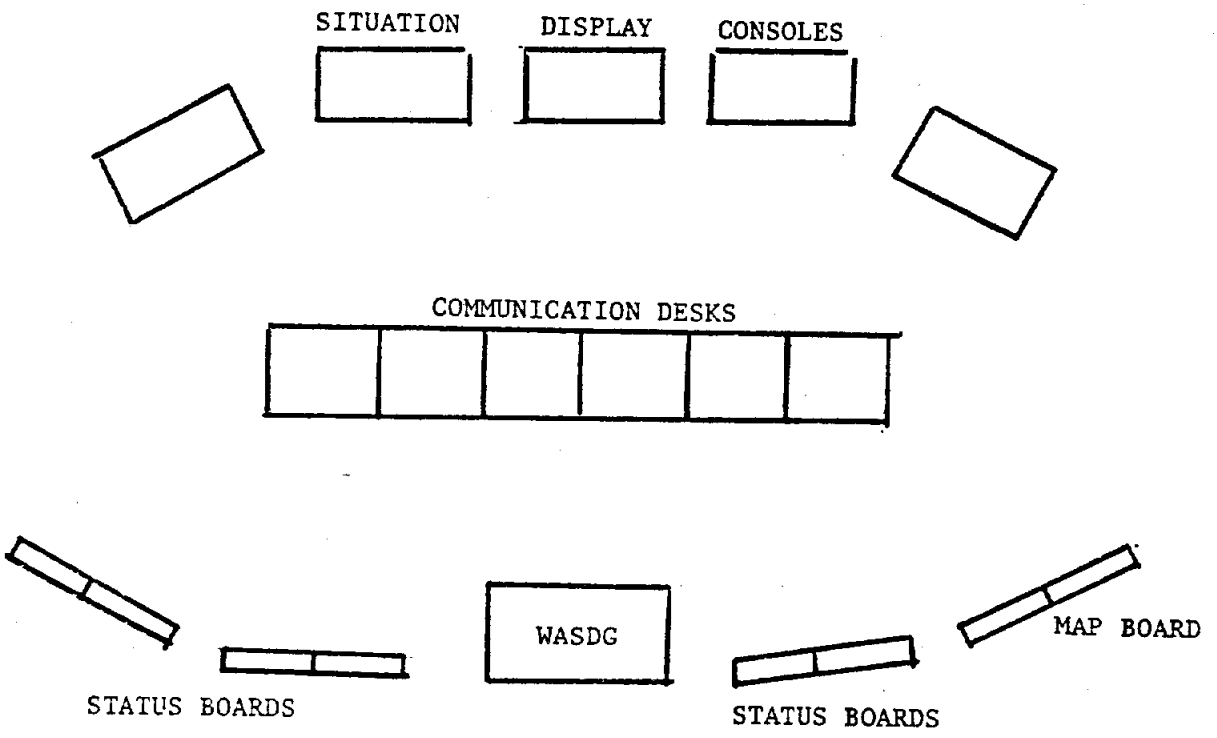


Figure 3-1. Equipment arrangement of Operations section personnel.

Criteria for Selection of Measures and Methods

The evaluator should apply the following criteria in selecting and developing his measures and measurement methods:

1. Objective. Ideally the measures employed should depend as little as possible on human judgment, because data collection in which the human is the measuring instrument inevitably involves some inaccuracy and inconsistency. Human data collection requires (a) recognition of the behavior to be reported (i.e., its occurrence), (b) determination of the relevant characteristics of that behavior, and (c) recording of the relevant data. Since considerable interpretation is involved in each of these phases, they are subject to error.

The previous paragraph began with the word "ideally." As a matter of practicality and cost, many measures employed in OST cannot be completely objective and, as will be seen subsequently, it would be undesirable to restrict measures to those that are completely objective.

2. Quantitative. Quantitative measures can be scaled and combined with other quantitative data; this is not true of qualitative data. Although qualitative measures can be useful to support or to explain the meaning of quantitative data, they cannot stand alone.

3. Unobtrusive. The act of gathering data should not affect the manner in which the operator performs his tasks. All data collection agencies (physical as well as human) should ideally be invisible to the performer. If operators become unduly aware of these agencies, they may perform in ways that are not representative of their operational activity. Operators may try too hard, they may make errors because of anxiety, and they may deliberately distort their behavior to fit some preconceived notion of what the evaluator wants. Since the goal of the OST is to verify operational performance, anything that leads to nonrepresentative behavior reduces test validity.

Test managers and operational commanders also generally object to any interference with ongoing tasks. An obtrusive data collection method (one which is highly visible) may be viewed as interfering, even if it does not hinder operational activity.

4. Easy to Collect. Any measure whose implementation makes extensive demands on the capability of data collectors is likely to produce errors in the data gathering process. If the data collector is required to make a difficult perceptual discrimination or computation, he may well make an error, thus reducing the validity of the resultant data. If, for example, he were required to note the precise time at which each signal on an air surveillance radar appeared, he would quickly lose track.

5. Require No Specialized Data Collection Techniques. There are several reasons why it is undesirable for data collection techniques to be highly specialized. Such techniques make it necessary to provide extensive training for the data collectors and, although some training should obviously be given them, it should (for time and cost reasons) be as little as necessary for them to perform their jobs. More important, special data collection techniques are likely to make it impossible to employ Marine Corps personnel as data collectors, because they will probably lack the needed technical background. It is always desirable to utilize operational personnel as data collectors (if data collectors are required) because (a) their familiarity with the task being evaluated may improve the precision of the data they collect, (b) operational personnel are less likely to be viewed by task performers as obtrusive elements, and (c) data collectors with the necessary technical background are likely to be scarce, and their unavailability may constrain data collection.

6. Require No Specialized Instrumentation. If the OST is being conducted in a field environment, specialized instrumentation may not function too well in that environment. Such instrumentation is often too delicate for the rough usage it may get in the field and therefore tends to malfunction more often than simpler equipment. Also, sophisticated instrumentation will require specialists to operate and maintain it.

7. Cost Little or Nothing. Cost is often the reason given by test managers for not conducting personnel performance tests. In most cases, this is only a rationale for rejecting procedures that these managers do not understand, but obviously specialized measures may require special instrumentation and personnel, and these may indeed be costly.

Criteria are of course ideals, and in the real OST world it is often impossible to satisfy these criteria completely. It may be necessary to use qualitative, subjective, or obtrusive measures or special data-gathering techniques, much as one would wish otherwise. The reason for listing these criteria is to provide a standard at which the evaluator can aim, rather than to make them hard and fast rules which often cannot be followed.

Available Measures and Methods

A listing of available objective and subjective performance measures and methods for use in evaluating both individuals or teams is provided on pages 3-7--3-9.[1] These measures and methods are discussed in the following pages organized according to the following outline:

1. Definition.
2. Information provided (by the method).
3. Use factors (factors to be considered in using this method).
4. Problems (in using this method).
5. Example.
6. Summary (of steps to be followed in using this method).

[1]Modified from a listing published by Smode, A. F., et al. The measurement of advanced flight vehicle crew proficiency in synthetic ground environments (MRL-TDR-62-2). Wright-Patterson Air Force Base: Behavioral Sciences Laboratory, February 1962.

LISTING OF AVAILABLE PERSONNEL PERFORMANCE MEASURES AND METHODS

OBJECTIVE

1. Time

a. Reaction time (i.e., time for an operator to perceive an event or start an action in response to some initiating stimulus).

b. Duration (i.e., the total time required for a task to be completed). In tracking a target, it is percent time on target.

c. Time between events (e.g., mean time between failures).

2. Accuracy

Accuracy in:

a. Observing and identifying stimuli or occurrences (internal or external to the system).

b. Estimating distance, direction, speed, time of movement of objects.

c. Detecting a change in events or stimuli over time.

d. Recognizing a signal in a noise or high target density background.

e. Recognizing an out-of-tolerance condition.

f. Positioning a control or a weapon or using a tool.

g. Reading displays.

h. Selecting among alternatives (e.g., decision making).

i. Making a series of discrete responses (e.g., throwing switches).

j. Making a series of continuing responses (e.g., tracking).

k. Communicating.

l. Error analysis, in terms of:

(1) Amplitude

(2) Frequency

(3) Type

(4) Changes over time

3. Frequency of occurrence

a. Number of responses made by operator in performing tasks such as the following:

(1) Observing.

(2) Controlling and manipulating.

(3) Communicating (e.g., verbal or written reports, requests for information, etc.).

(4) Personnel interactions.

(5) Maintenance diagnostic checks.

b. Number of effects of task performance, such as the following:

(1) Number of errors.

(2) Number of out-of-tolerance conditions.

4. Amount Achieved

a. Cumulative responses (i.e., total number of responses made).

b. Degree of success in accomplishing tasks and mission.

c. Achieved reliability (percent of tasks accomplished, or ratio of tasks successfully accomplished to all tasks undertaken).

d. Proficiency scores (e.g., written test performance).

e. Terminal or steady state value (e.g., a temperature high point).

f. Performance variability (e.g., degrees change per hour).

5. Amount Consumed

a. Physical resources consumed in terms of activity or time, such as:

(1) Fuel/energy consumed or conserved.

(2) Units consumed in accomplishing tasks (e.g., weapons or ammunition expended).

b. Personnel resources consumed in terms of casualties.

c. Man-hours consumed (response time x number of personnel involved).

SUBJECTIVE

Available subjective measures are:

1. Interviews.

2. Questionnaires.

3. Observations.

4. Ratings.

5. Checklists.

6. Notation of critical incidents.

These measures are used for the following purposes:

1. Performance Efficiency Determination.

a. Performance of the operator/team (ratings).

b. Success of the task/mission (ratings).

c. Adequacy of equipment, procedures, logistics, technical data, and training (checklist, interview, ratings).

2. Factors Affecting Performance.

a. Identification of factors affecting performance (interview, questionnaire, ratings).

b. Attributes possessed by task and performance (e.g., degree of comfort) (ratings).

3. Event Occurrence.

a. Description of task performance (observation, interview).

b. Unusual occurrences (critical incidents).

In the previous section, we said that an operational system evaluation has the following goals (as these relate to personnel performance):

1. Verification that the system (including its personnel) satisfies system requirements.

2. Prediction of the future operational performance of the prototype system under evaluation.

3. Diagnosis of problems found as a result of testing the system.

4. Provision of data that may be applied to the analysis of current and future systems.

5. Identification of the personnel resources needed to utilize the system effectively.

6. Provision of data for future training of personnel who will operate the system being evaluated.

Table 3-1 categorizes each measure/method listed above in terms of its potential use in the evaluation for meeting these goals.

Objective Measures and Methods

Time

Reaction Time.

1. Definition. Reaction time (RT) is the time between the occurrence of an event requiring an action on the part of the operator or team (the initiating stimulus) and the start of the action (the RT response) demanded by the event (see Figure 3-2). Since the initiating stimulus must be recognized by an operator, that stimulus is likely to be something observed directly by him or displayed on a CRT or indicator display. The RT response is the operator's action in throwing a switch, making a verbal report, etc.

2. Information Provided. Table 3-1 indicates that the RT measures can be used to verify (or fail to verify) that personnel can meet system requirements and that the data can help to predict the operational performance of the new system as well as to provide data useful in planning future systems of the same type.

The major purpose of measuring RT is to determine how quickly the operator/team can react to an initiating stimulus. Before selecting this method, the test planner must ask whether this information is necessary for the evaluation.

Where the operator or team is required to respond to an initiating stimulus in a fixed (minimal) period of time, his or its ability to do so must be verified. If the RT required of an operator is very short, he may have difficulty accomplishing the task. If one examines task requirements in advance of evaluation, and required RT appears to be quite short, then RT measurement may determine if the operator's capabilities are being exceeded.

If the operator/team is required to react as quickly as possible (even though a fixed RT is not specified), the test planner will wish to determine the minimum and maximum RT he can expect of the system. However, this information cannot be used to evaluate the effectiveness of the system unless an RT criterion (standard) exists or can be developed. If an RT criterion is desirable but does not exist, the operational personnel who must operate the system can develop such a criterion by the method described in Section Two. If a time requirement does not exist, either explicitly or by implication, RT is unlikely to be of value in evaluating the system.

The nature of the system will determine whether or not an RT measure is meaningful. RT is significant only when the speed of a reaction to the initiating stimulus will determine the effectiveness of the system response.

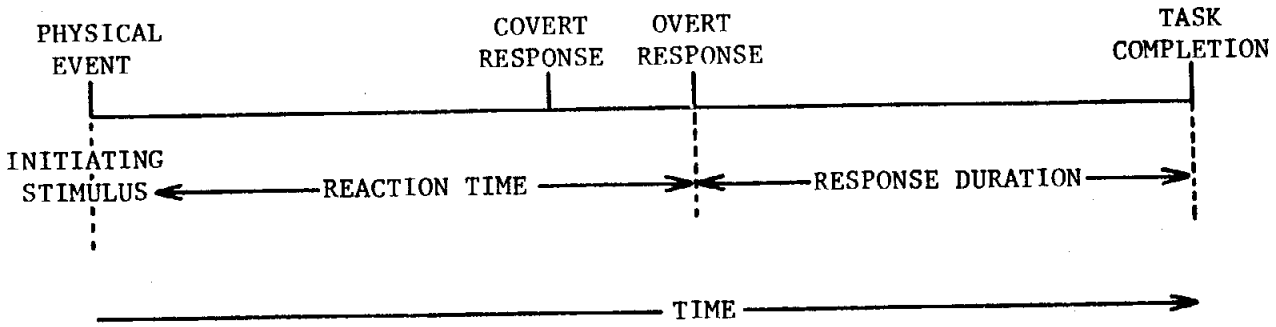


Figure 3-2. Reaction time and response duration measures.

Table 3-1

Information Provided by Measures and Methods

Use	Objective					Subjective					
	Time[a] Measures	Accuracy	Frequency	Amount Achieved	Amount Consumed	Inter-views	Question-naires	Observa-tions	Ratings	Check-lists	Critical Incidents
Verification of System Performance	X	X	X	X	X			X	X	X	
Prediction of Operational Performance	X	X	X	X	X			X	X		
Diagnosis of Problems	X	X				X	X	X		X	X
Data for Future Use	X	X	X	X	X			X	X		
Identify Personnel Resources Required						X	X	X			X
Data for Future Training		X				X	X	X	X		X

[a]Reaction time, duration, and time between events.

Beyond the verification that personnel requirements have been satisfied, RT can be useful in suggesting the speed with which the new system, when deployed, will respond. This obviously impacts upon its ability to accomplish its operational goals. Moreover, if a new advanced system of the same type is to be procured in the future, the RT data will supply information with which the performance of the new system can be compared.

3. Use Factors. A distinction must be made between operator and team RT. The team RT may be distinctly different from that of the individual. An officer in charge of an artillery unit may respond to the sight of an enemy target (the initiating stimulus) by issuing a command to his battery to fire (the officer's RT response). The officer's command then becomes the initiating stimulus for the battery (team) and the firing of the first round becomes the team's RT response.

In many cases, the determination of operator RT may be much less important than team RT. The individual operator's RT is, in many cases, almost immediate and thus may have minimal effect on system RT (except where the operator is in effect the system, as in a single-pilot aircraft). Team RT may be much longer, however, because of team member interactions that delay the team response and may have significant impact on system RT.

Once the team has been presented with the initiating stimulus, the RT of any individual in that team need not be measured. To the extent that team RT depends on the individual RTs of its members, the latter is included in the former. Whether one measures individual or team RT depends on which is being evaluated. If the team RT is excessive (i.e., greater than that allowed by the criterion or whatever operational commanders feel is excessive), it may be necessary to measure the RTs of individual team members to determine the source of the difficulty.

If RT is important, it must be measured precisely. Where short time intervals are involved (fractions of seconds, although in most cases no shorter than seconds or minutes), RT measurement may require the use of highly sophisticated instrumentation, particularly in the case of operator RT. Team RT is usually somewhat longer and may not require such precision; a stop watch may be sufficient.

In collecting RT data, a preliminary analysis should be made of all the operations involved in the task to be measured. This means plotting out each step in the task in terms of the time it should take to perform the step, including required RTs to initiating stimuli.

4. Problems. The recognition of the initiating stimulus may be difficult. Where the stimulus cannot be easily anticipated by an observer, when it is accompanied by irrelevant stimuli from which the observer must distinguish it, or when the initiating stimulus is itself very short, it may be difficult for both the operator and the RT data collector to recognize that stimulus. Suppose, for example, that one wished to determine the operator's RT to the appearance of an enemy aircraft on a radar scope. It is possible for the data collector to note when the echo first appeared on the scope. However, the observer's RT to the appearance of this stimulus is much the

same as that of the operator whose performance he is measuring; that is, it is compounded of the time needed to detect the blip, to distinguish it from clutter, to note whether the operator has detected it, to throw the switch on the timer, etc. This means that the observer's error in recording this form of RT will be equal to the RT for the operator. This type of situation demands such precise time recording that often only special instrumentation has the needed capability.

Another problem is the nature of the RT response. In situations in which the operator's response is covert (i.e., perceptual or analytic and not expressed as a control manipulation), it is extremely difficult for an observer (or for that matter, for instrumentation) to pinpoint exactly when the response has occurred. In surveillance systems, the operator may initially respond to the appearance of a target by continuing to observe it. Unless he is required to report verbally or to throw a switch indicating that he has recognized the target, there will be no observable indication that he has in fact recognized the target. In situations involving perception, monitoring, silent analysis, etc., there is no overt response. In such cases, it becomes necessary to wait until the operator activates a control and to consider this activation as the RT response.

5. Example. A major function of the TACC is to control air space over a specified area of responsibility. This requires the Air Defense Coordinator (ADC), for example, to detect and identify on his CRT any aircraft entering his air space. Since the aircraft may be unfriendly, the more rapid the detection and identification, the more efficient the system.

The appearance of a blip on the CRT is the initiating stimulus for the ADC. His RT response is his perception (recognition) of the unidentified aircraft. Since this recognition is covert (i.e., not observable or measurable without extremely elaborate instrumentation), a more feasible RT response would be his "hooking" the aircraft (a control action) to secure additional information about it.

The most important question in the selection of an RT measure for TACC evaluation is whether such a measure is meaningful in the TACC context. Performance standards for TACC operations do not include a quantitative RT requirement although, obviously, RT should be as short as possible. Since there are five operators controlling various aspects of the air space, it would seem reasonable that, if RT were collected for one operator, it should be collected for all. However, one might reduce the magnitude of this task by selecting only critical personnel ("key men") and measuring only their critical functions.

Moreover, the operator performs several sequential functions for which an RT measure should be secured if RT is to be measured at all. For example, the ADC may have to make an unscheduled launch of an aircraft to intercept an unfriendly aircraft and, obviously, the time between identifying an unfriendly and launching an interceptor is critical.

With a system the size of TACC, one can select many RT measures. If the TACC had the capability of automatically recording the time at which

each control was activated and/or of recording verbal communication on a time base, the problem of measuring RT would be much reduced, although the data analysis of the automatic printouts would still be considerable. (Such an automatic monitoring system is in use aboard ship with the Navy Tactical Data System.)

The TACC represents an evaluation situation in which RT measurements would be highly desirable but in which the complexity of the situation, without automatic instrumentation, makes this measurement difficult. The absence of a performance criterion makes interpretation of the resultant data also problematical, because what would any particular RT value mean? If it takes the average operator 6.8 seconds from the first appearance of a target on the screen to "hooking" it, is this good or bad? Again we see that the critical factor in this measurement, as in other measurements, is the existence of a criterion.

However, if it is unpromising to attempt to measure the RT of individual operators, is it possible to determine a total TACC RT? No, because each of the operators performs different functions concurrently, so that system activity is a composite of many actions which cannot be meaningfully combined, as they must be if one attempts to secure system RT.

6. Summary. The following steps should be followed in selecting an RT measure:

a. Analyze system requirements to determine whether an RT measure is needed, will be useful, and is feasible in evaluating personnel performance.

b. If the answer to item (a) is yes, analyze task operations to determine:

(1) Whether individual or team RT is needed.
(2) What the initiating stimulus is.
(3) What the RT response is.
(4) The measurement precision required.

c. Determine whether RT can be measured by observation (data collector) or requires instrumentation.

d. If RT is to be measured by observation, ensure that the data collector is aware of what the initiating stimulus and the RT response are and require him to practice the RT measurement.

Duration.

1. Definition. Duration is the time spent accomplishing a task, from the time task performance is triggered (the initiating stimulus in RT measurement) to the time the task is completed (Figure 3-2). This type of measure is an extemely common one. It is often recorded even when there is no apparent need to verify a duration requirement.

2. Information Provided. Duration is important when the system prescribes a maximum duration for a task or group of tasks. If a task must

be performed in no more than 5 minutes, for example, it is important to measure task duration to determine if that requirement is satisfied. The question is one of operator/team capabilities: Can the task(s) be performed in no more than 5 minutes, or does this requirement impose a demand on the operator/team that is physically not possible or, as a consequence of which, accuracy is degraded? An example of critical task duration is often found in maintenance operations where a maximum time for restoring a malfunctioning equipment has been assigned.

Even if a maximum duration is not specified, it is often of interest to the test planner to determine performance duration: a performance duration that is considered excessive by operational commanders will degrade system effectiveness. If, during the OST, difficulties arise in completing a function, duration measurements could be useful in suggesting the cause of the difficulty: a particular subfunction which takes excessively long may be the factor causing the difficulty. Duration data also are useful in predicting the operational performance of the new system and in providing a basis for comparing the performance of future systems with the present one.

If system operations are flexible (as in the TACC) and the operator is performing several functions concurrently or sequentially, the test planner may wish to determine the percentage of his time the operator spends on individual functions (or the time spent performing the task and the time occupied by equipment operations). If an operator, for example, has to receive telephoned messages, to record them, to enter the data into a computer system, etc., the percentage of time he spends on each of these functions may indicate if he is being overloaded by a particular activity.

3. Use Factors. Ordinarily this type of time measurement does not have to be extremely precise, unless there is some system requirement which necessitates highly precise measurement (e.g., as in timing a foot race). When duration is relatively gross, it is easily recorded. For very precise measurements, instrumentation may be required.

Because most duration measurements usually need not be overly precise, the measurement operation is comparatively simple, often demanding no more than a stop watch and a single observer. Where several concurrent tasks must be individually measured, however, several data collectors may be required.

Measurement of team task duration should pose no more problem than measurement of individual task duration. A team task is one in which the individual functions interact to secure a common output. If, as in the case of TACC, distinctly different functions are performed by individual operators, one cannot speak of the total number of these functions as a team task.

The scope of the team effort whose duration is to be measured will obviously be greater than that of any individual task, and it will be necessary prior to measurement to define that scope. As in the case of RT, the duration of the individual tasks comprising the team effort need not be measured unless duration either is or is expected to be excessive.

4. Problems. One possible problem is the precise definition of the unit of performance whose duration is being measured. If duration is measured from the start of system operation to its conclusion (when the system is secured or goes "off the air"), it is relatively easy to record. The measurement of individual task duration (i.e., one task nested within a series of tasks or several tasks performed concurrently) is more difficult. Tasks in a series flow into each other and there may be no clearly defined start and stop point to bound the limits of the measurement. This is another reason for analyzing task characteristics prior to measurement.

The duration of system operations may not be the same as the duration of tasks performed during those operations. Tasks may be performed prior to the equipment taking over, or the operator may have to wait while the hardware is performing its part of the operation. Measurement of the system operation as a whole will therefore include the task, but will not necessarily spell out how long the task took.

5. Example. Once TACC is set in operation, it functions on a continuing 24-hour basis. System duration for TACC as a whole would therefore not be very meaningful. On an individual operator and function basis it would, however, be quite feasible to determine duration. One might, for example, determine the length of time it takes to intercept an unfriendly aircraft; the duration would start from the launch of an interceptor to the time the interceptor is within firing range. From an evaluation standpoint, this datum might be of little value, but it might be significant for diagnosing the decision-making efficiency of the operator, since intercept duration is partially determined by the vector he gives the interceptor pilot.

Since TACC personnel perform many functions, it may be of some interest to determine the distribution of time among these functions (e.g., the percentage of time spent coordinating resources as against the time spent launching aircraft). This information, too, would be primarily of diagnostic value.

6. Summary. The following steps should be followed in measuring duration:

a. Determine whether a duration measure is needed.

b. Determine the unit of performance whose duration will be measured.

c. Determine whether duration should be measured by instrumentation or manually.

d. If duration is to be measured by observation, ensure that the data collector is aware of the start and stop points of his measurement.

Accuracy.

1. Definition. Accuracy, or its converse, error, is probably the most common and perhaps the most useful measure of personnel performance.

There are systems and tasks in which reaction time and duration are not important but accuracy is critical in all. All personnel performance assessment depends wholly or in part on accuracy measures.

Except in those rarely found systems that permit no error (and it is hard to think of an illustration), one would expect some errors to occur. The problem then becomes one of evaluating the significance of the number and type of errors to system performance. Each error is not equivalent to every other error. Even if it were, the relationship of a particular number of errors in performance to the system's capability to accomplish its functions depends on a criterion of the maximum number of permissible errors. For example, suppose that in a new aircraft crewmen made an average of .75 errors in performing a preflight checkout of 38 steps. Does this mean that the checkout procedure for this aircraft was ineffective or that personnel could not perform preflight checkout as required? One cannot answer this question unless the evaluation begins with a standard that N errors of a given type are or are not acceptable. The figure of .75 errors is, therefore, just an interesting statistic.

Any inadequacy in performance can be considered an error, but errors generally consist of incorrect commission or ommission. Errors of commission include performing a nonrequired action and performing a required action incorrectly or out of its required sequence. Errors of omission consist of failing to perform a required action. Errors may occur when personnel do one or more of the following:

a. Utilize data from their equipment (e.g., reading a temperature gauge incorrectly or failing to read it when required).

b. Fail to take action demanded by the equipment (e.g., failing to take emergency action when meters indicate the need to do so or failing to report a potential malfunction).

c. Utilize the environment or objects in the environment (e.g., failing to recognize a geographical landmark or an enemy target).

d. Fail to take action demanded by data describing other systems (e.g., failing to comprehend intelligence on enemy movements).

e. Estimate distance, direction, speed, time (of enemy movements) or make projections of action required on one's own equipment or personnel in the future.

2. Information Provided. Accuracy data are critical to verify that personnel can perform required tasks. Obviously, an inability to perform adequately is manifested by error. To the extent that the system performs during OST as it will perform operationally, accuracy data serve to predict operational performance. Accuracy data assist in the diagnosis of problem areas that arise during OST. The nature of the error, who makes it and when, may reveal an inadequacy in system design that impacts on performance. Accuracy data suggest where training emphasis should be placed, since errors reveal functions that must be strengthened through additional or different

training. Accuracy data are also applicable to future systems since, like RT and duration, these serve as a baseline against which to compare the performance of more advanced systems.

3. Use Factors. It is necessary to know not only error frequency, but also the type of error made. The nature of the error is potentially diagnostic of a situation that needs modification; it may cast light on what needs to be done.

Should an error that is corrected by the operator be counted in the same way as one that is uncorrected? No. Many more errors are corrected than are not; if this were not so, it is unlikely that any equipment could be operated effectively. An error corrected by the operator would not be counted unless it degrades task performance significantly until corrected.

The evaluator also needs to know the criticality of the error. Some errors have potentially significant or catastrophic effects on task and function accomplishment; others do not. For example, an error in performing a continuous function such as tracking may be much more significant for performance than one in a discrete task like throwing a switch, where the error is more visible and can more readily be reversed.

In general, errors that have only a minor effect on performance and that are readily corrected need not be counted in the determination of personnel effectiveness. The more significant the potential effect of an error, the greater weight it should assume in the evaluation. It is possible to weigh errors on a criticality scale such as the one below.

Rating	The error has	As a result
4	Extremely grave effects.	Task cannot be accomplished.
3	Great effect on performance.	Task is accomplished, but inadequately.
2	Moderate effect on performance.	Task is accomplished adequately but is markedly delayed.
1	Slight effect on performance.	Task is accomplished and with only slight delay.
0	No effect on performance.	Task is accomplished adequately and with no delay.

Assume one error rated 3, one rated 1, and three rated 0. The total error score assigned to a particular performance would be 4. Such a technique permits one to compare performances on different tasks. However, it does not provide an absolute evaluation (e.g., the probability of this task being performed correctly in future system operations is .75).

In evaluating the potential consequences of an error, the evaluator must also differentiate between the effect of the error on the individual task and its effect on overall system performance (where more than one task is involved). Some errors impede the performance of the individual task but have little effect on overall system performance, either because the task is not critical to that system performance or because there are compensatory mechanisms in the system which cancel out the error effect. Obviously, a task which, if performed incorrectly, would seriously jeopardize system mission performance would be much more critical than one which affected only the individual job.

The criterion of performance accuracy (which should be defined prior to measurements) should therefore indicate not only the number of allowable errors, but also the type of error which is to be considered in making the assessment and its criticality. Knowing the types of errors that can be made will make it easier for the evaluator to recognize when that error occurs in observing performance. Without such a prior analysis, it may be difficult for the data collector to recognize performance inadequacies. Of course, prior experience with a similar task or system may compensate partially for lack of such an analysis.

Errors can be recorded manually or by instrumentation. Instrumentation is necessary when the erroneous responses: (a) are numerous or frequent, (b) change frequently, (c) have a very short duration, or (d) are difficult to observe directly (e.g., physiological reactions). Manual error recording (by observation or self-report by the operator whose performance is being evaluated) is necessary when (a) considerable interpretation of the response is required in order to determine that an error has in fact occurred, or (b) the error response is recognizable only by the context in which it occurs (machines have difficulty recognizing context). Errors may be recorded manually when (a) operator responses are not overly many or frequent, (b) have a reasonable duration, or (c) are readily recognizable. Statements such as "Not many or frequently" are rather general; their specific values depend on the individual data collection situation. However, even slight experience with the operator responses to be recorded will permit the test planner to make a decision on these points. The tradeoff between automatic and manual data recording of accuracy data is often determined by cost. Test managers prefer not to spend money on instrumentation if they do not have to.

The test planner almost always has a choice about the data he can record. Most systems involve various types and levels of personnel responses and the test planner may elect to record only some of these. Because a choice is possible, it is essential that the test planner specify in advance of the evaluation precisely the kind of data he will collect. Whether data are collected manually or by means of instrumentation, the operator's response must be precisely defined. If it is not, it may be difficult for the data collector to recognize the occurrence of the error.

The previous section recommended that the test planner record measures on those task levels appropriate to the evaluation questions to be answered (e.g., recording the number of target hits rather than a rifleman's finger-squeeze pressure). Within that general constraint, as much error data as

possible should be recorded because varying conclusions are possible depending on the particular errors made.

Should all errors be recorded? The answer is no, and this ties into the preceding discussion of error criticality. Errors that can have no impact on system performance can be ignored (assuming that they are recognized during data collection as having no impact); all other errors should, of course, be recorded. Theoretically, if all personnel actions are relevant to system performance, all errors should be significant. However, in actual practice, it often turns out that certain of these are trivial. The preevaluation analysis of the data to be collected should include, along with a review of the operator's task and the kinds of performances that may occur, an analysis of the effects of these performances.

The determination of what constitutes an error presents special problems when the operator's performance is covert. (See the previous discussion on reaction time.) By this we mean that the task is being performed inside the operator's head. Such tasks are likely to be perceptual (e.g., monitoring a display of some sort) or cognitive (e.g., making a decision or a mental calculation). In such cases, the test planner should look for some overt response which is associated with the covert one. For example, if the operator must throw a switch or make a verbal report based on his covert activities, it is throwing the switch or reporting verbally that become the types of data whose accuracy is recorded. The preevaluation analysis of data to be collected should indicate both covert and overt responses.

Manual techniques for recording errors include the following:

a. If all (or most) potential errors can be categorized in advance of observation, one could develop a checklist of these errors and simply check them off as they occur during task performance. However, this procedure is not recommended for two reasons: First, it is difficult to anticipate all errors. Second, simply noting that an error of a given type has occurred does not supply all the information desired. One wishes to know also at what step in task performance the error has been made and what the operator does (or does not do) as a consequence of his making the error.

b. One could attempt to record all of the task actions exhibited by the operator and later sort out those actions that are erroneous. However, such complete data recording is difficult for an unaided observer; he would need something like a videotape camera. The advantage of comprehensive performance recording (such as one could have from a filmed record) is that it provides a complete record for later analysis, permitting the evaluator to examine the performance more leisurely. This procedure is not recommended, however, unless instrumentation is available.

c. The most common procedure is to use an operating procedure as a sort of "template." As the operator performs his task, the data collector checks off each action on the procedure. A performance deviating from that specified in the procedure would be noted at the appropriate step in the procedure (Figure 3-3).

Task Step	Error and Consequence
1. Turn power switch to ON position.	
2. Read external temperature.	
3. Align power output reading as required by external temperature.	
4. Check CRT gain intensity by depressing GAIN switch and read value displayed on GAIN display.	
5. Record gain value.	

Figure 3-3. Use of operating procedure as error data collection device.

Difficulties arise, however, where the task to be performed is so unstructured (i.e., involves so many contingencies) that a step-by-step procedure for that task cannot be developed. Although one could still use an operating procedure, it would be more general than the one shown in Figure 3-3. Instead of listing discrete steps, the test planner could list the general functions to be performed (e.g., launch aircraft, coordinate assault) and, below these, the major actions required to perform each function. Thus, in vectoring an interceptor to the vicinity of an unfriendly aircraft, the operator has to communicate with the interceptor and make a number of course corrections. Any errors associated with each of these actions would be recorded. Naturally, the more unstructured the task, the more the observer/data collector must know about the job being performed so that he can recognize an error occurrence.

As was pointed out previously, analysis of the errors made, their frequency and amplitude, and where in system operations they occur suggests the potential causes of the error. The type of error is important, since it often reflects the cause.

Basically, however, the evaluator wishes to determine the efficiency of operator performance. This can be done by contrasting the number of tasks successfully accomplished (s) as against the number attempted (n). The ratio (s/n) provides a percentage value which can be interpreted to indicate the future likelihood of these tasks being successfully accomplished. For example, if 10 tasks are attempted (n) and 9 are correctly performed (s), the efficiency ratio if .90. The emphasis here is on successful accomplishment of a task; if errors are made but are retrieved (or even if not retrieved) and the task is completed, it must be recorded (despite errors) as successful. This relates to the criterion of what constitutes an error. One is not simply counting errors; one is trying to determine whether the task was successfully performed.

The determination of successful task accomplishment is essentially a judgment made by the observer/data collector. That judgment is based on the output of the task performance when it is completed. For example, an air intercept is judged as successful when the interceptor is vectored to a point in space where the pilot can report viewing the unfriendly aircraft.

4. Problems. Many of the problems encountered in applying accuracy measures have been discussed as part of the preceding section.

Special problems may arise, however, in measuring team accuracy as distinct from individual operator accuracy. This can be done without excessive difficulty provided there is a product of the team activity. Assume a four-man team in which a decision (e.g., to reconnoitre a particular area) must be based on inputs from three of the men (the fourth man being the decision maker). The accuracy of that decision can obviously be measured. The decision is a team output (not merely the result of the decision maker's activity) because the decision could not be made without the three inputs.

One would not wish, however, merely to examine the team output (decision) without considering the individual performances that led to it. If the decisions were invariably correct, then perhaps one could ignore the individual performance, assuming that they were correct because the output was correct. However, decisions or other team outputs are sometimes incorrect; when this occurs, it is necessary to examine the individual inputs that led to the team output. From the standpoint of the accuracy or quality of the team output, it is essential to measure individual contributions as well as the team output.

A distinction must also be made between an erroneous output and an error made in arriving at that output. As has been pointed out previously, many errors are inconsequential. However, an erroneous output is always significant since it directly impacts on the mission of the team. Hence, a team output measure (when available) should always be gathered.

5. Example. One of the factors that will affect the collection of accuracy data in the TACC is that there is a variable workload (e.g., the need to detect and identify aircraft and to launch interceptors, depending upon enemy incursions). One would therefore expect TACC personnel performance to vary with that load. Thus, TACC performance should be sampled at various times to record performance as a function of the range of loads from very light to very heavy. If the evaluation is performed using "canned" stimuli (e.g., targets produced by a target generator), then that load can be specified in advance. If the evaluation is not performed according to a prepared scenario of inputs to the system, then data collection will depend on targets of opportunity. In either case, since workload for the TACC does ordinarily vary, data sampling is necessary.

The number of activities performed in TACC lends itself to a variety of error measures. There is no question that accuracy measures should be taken, but which ones and how? There is the rather specific procedure employed to operate the situation display consoles--such things as depressing

the correct buttons to "hook" the target or to select a piece of video, etc. The number of errors made in operating the situation display console could be determined without difficulty by observation of the operator's activities. The data collector would have to know its operation since the procedure is so flexible (depending upon the nature of the inputs) that it is impossible to follow a written procedure exactly.

Other operator activities are more difficult to record. Analysis of the functions of TACC personnel reveals that there is considerable decision making, supervision, and coordination. Part of the time they are monitoring the buildup of new tactical situations; much of this activity is covert or, if not covert, at least it cannot be tracked with the aid of a step-by-step procedure. Under these circumstances, it will probably be necessary to make use of ratings of performance quality, either by supervisors or by a knowledgable observer. A more detailed discussion of the form such ratings could take will be given in the section on ratings.

The great variety of activities performed by 11 or more personnel requires some degree of selectivity on the data collector. Even if it were feasible to have one observer for each TACC team member (and that is unlikely) when input loads are high, data collectors will not be able to keep up with the flow of activity. It therefore is necessary to preselect those "key" activities which are most representative of and critical to TACC operations and to concentrate on these. Hence, prior analysis of the functions and tasks performed will be necessary. This analysis should indicate also the kinds of errors that are most likely to occur; this information will help to cue the observer to what he should look for.

Another consequence of the size of the TACC team and the variety of their activities is that instrumentation is not a very feasible way of collecting error data. The covert nature of TACC activities such as coordination forbids their instrumented measurement. To replace an observer, one would need at least three or four video cameras and even these would not provide the degree of detail needed, because it would be difficult to get the cameras close enough to record adequately without their becoming intrusive. And then, of course, the posttest analysis of this material would be a horrendous task. This is a case where the flexibility of an observer, assuming he is properly trained for the data collection task, makes him much more valuable than instrumentation.

6. Summary. The following steps should be followed in measuring accuracy:

a. Prior to collecting error data, perform an analysis to determine the following:

(1) The criterion of acceptable accuracy.

(2) The kinds of errors that may be made and those that will be recorded.

(3) The effect of the error on the individual task and the overall system.

(4) Whether errors should be recorded manually or by instrumentation.

b. In analyzing error data, weigh the error by its effect on the task.

c. Where possible, use an operating procedure as a guide to the recognition of error.

d. Where possible, determine the number of tasks successfully accomplished and apply the s/n ratio.

e. In a team situation, measure the accuracy of the team output and the individual contributions of team members to that output.

Frequency of Occurrence

1. Definition. The test planner may also wish to determine how frequently the operator's responses occur. Frequency, which is occurrence as a function of some time interval, is simply the tabulation of personnel actions (or events occurring as a result of personnel actions) as a function of time or other events occurring during system operations. In the latter case one might, for example, record frequency of error as a function of several stages of an operation. Frequency is easy to secure, provided one can sort personnel actions on a time base.

2. Information Provided. Occasionally, a standard of personnel performance will specify a required frequency of response (in which case it is necessary to verify that personnel do respond with that frequency), but this does not happen too often. Frequency data are of primary use in a supporting role; that is, to help explain other data more directly descriptive of personnel performance.

This information may illuminate factors that have affected the operator's performance. The relative frequency of certain types of error, for example, may suggest special difficulties the operator has in using his equipment. For example, if he makes a disproportionate number of errors in operating a particular control, that control may have been poorly designed or it may have been located in an awkward position. In evaluating the efficiency of a squad, the frequency of certain types of verbal reports from one squad member to another may indicate the degree of squad coordination.

3. Use Factors. Any series of discrete operator actions can be recorded as a frequency. Ths most common such applications are:

a. Types of personnel actions performed.

b. Types of error made in operating an equipment.

c. Types of verbal reports communicated from one action station to another.

d. Types of maintenance actions.

4. Problems. As with any other type of data, the actions whose frequency is being tabulated must be observable and distinguishable from other actions. Given this, it is relatively easy to secure frequency data. It is, however, much harder to interpret its significance, since frequency has only one dimension (occurrence) and does not by itself explain anything.

5. Example. Since many different functions (e.g., monitoring CRTs, communicating, analyzing map and status boards, etc.) are performed in TACC, it would be of interest to determine the distribution of these activities over time (an 8-hour shift, perhaps). Such data might have diagnostic value in terms of indicating where personnel are overloaded. For example, if at certain times one operator is burdened with a high frequency of telephone messages, it might be desirable to supply him with a backup communicator. Since, however, there is usually no system requirement that enjoins a particular frequency of activity, frequency data would be collected only if it did not interfere with other data collection requirements.

Frequency data can be secured in TACC at a number of levels. The most molecular would be the frequency with which a situation display operator uses certain controls in monitoring his assigned airspace (e.g., hooking, shifting sector scan, etc.). It is difficult to see the special utility of collecting this type of information, however, because, at this level, control use frequency is determined by the number and type of aircraft entering the airspace. Moreover, unless the data collection procedure was automated, it would tie up one data collector for each situation display operator being monitored.

At a higher level, the performance frequency of the functions (described at the start of this example) could be determined. This could produce useful if not critical information. Frequency at this level would also require direct observation by data collectors, although it would probably not require a data collector for each operator. This type of frequency information would be facilitated by use of a checksheet, enumerating the various activities to be performed; the data collector could simply check off each activity as he observed it occurring. Naturally, before developing this checklist, the test planner would have to analyze the types of activities that could occur. This analysis also presumes the observer's ability to distinguish one activity from another. If the activities did not appear with high frequency and did not overlap, the observer could also note the time of their occurrence and/or their duration.

The types of errors made could also be noted as they occurred. Assuming these do not occur too frequently, they could be noted as part of the rating evaluation of accuracy noted in the previous section. Frequency of error type has considerable diagnostic value, since error type is logically related to some problem in aptitude, training, procedures, or input load.

Determination of types of verbal reports would best be determined from analysis of automatically recorded communications, since verbal communication is usually quite rapid and the data collector may not have time to categorize the message. Communication frequency has some diagnostic value, but not as great as error type. Communications analysis has value in terms of suggesting the efficiency with which the team as a whole interacts among its members and with other organizations.

6. Summary. The following steps should be performed before measuring frequency of occurrence:

a. Define the performances whose frequency is to be measured.

b. Determine the time base (e.g., hours/minutes, events, stage of operations) to which frequency is to be related.

Amount Achieved/Amount Consumed

1. Definition. Some military systems may require that a specified quantity of a product be output by the operator and/or system. In other cases, the test planner may wish to determine what the operator/system is capable of outputting, even though no specific criterion requirement exists. The measure utilized for this purpose can be considered the amount achieved.

The mirror image of amount achieved is consumption or the quantity used, either absolutely or per unit of time; something must be used up which is related to the goal of system operation. The criterion (either explicitly or implicitly) is to use as few resources as possible. On a system level (e.g., artillery battery or tank), the number of shells expended, the number of missiles fired, the amount of fuel burned, etc. are sample measures of consumption.

2. Use Factors. In order to apply amount achieved, the operator/system must output a product which can be mathematically treated. The most obvious civilian example of such a product is the assembly line, where number of units being assembled, installed, welded, etc. can be counted to secure a total value. Military systems, too, have products that can be counted. For example, the number of shells being moved by a transportation company can be reported as an amount achieved.

One measure of amount achieved was referred to in the section on Accuracy. This is the percentage of tasks successfully acccomplished (however one defines successful accomplishment). The measure is simply the ratio between tasks successfully accomplished (s) and all tasks attempted (n). Even for those systems in which there is no overt product, one can consider the accomplishment of a task as a product which can be quantified by the s/n ratio. If one can observe the operator performing and say that a task has or has not been successfully accomplished, one can apply this measure. However, the tasks must be discrete and there must be a concrete criterion of successful completion. This criterion need not itself be quantitative, as long as the evaluator can judgmentally determine task accomplishment.

The s/n ratio is, however, purely a binary measure and does not allow for degrees of success; in that respect, it is a rather crude measure. Moreover, to apply the ratio, any task being evaluated must be repeated a sufficient number of times. If the operator performs a particular task only once or twice, he may either succeed completely (thus giving him a performance value of 1.0) or fail totally (s/n ratio of .00). Either of these values is obviously a gross over- or underestimation of his true performance capability.

If the system has a specific quantifiable goal (e.g., hitting target X ± 100 yards), then a deviation measure (the extent to which rounds exceed allowable tolerance) can be utilized as a measure of accomplishment. The number of rounds on target or the time on target are also measures of amount achieved.

In most cases, system requirements will not specify the amount (of any product) to be achieved. Hence, the usefulness of this measure is primarily in describing how the operator/system performs and not to verify a given level of performance.

3. Problems. It is not easy to measure personnel performance achievement, because the product of personnel performance is often not concrete, as for example, in monitoring, planning, or deciding. Any overt personnel response can be counted; but even so, such a measure might have little relevance to the goal of the total task. For example, one could count the number of switches thrown by a console operator, but this number might have little bearing on what the operator is attempting to do, which might be to vector aircraft. From a personnel performance standpoint, consumption also is not an easy measure to apply since the operator often expends nothing of his own except his energy.

If a team has a concrete output, it can be measured in terms of accomplishment just as the individual operator's output. For example, consider an air defense center with six operators, each of whom is responsible in his own sector for detecting, tracking, and assigning weapons against enemy fighters and bombers. Each operator's performance can be measured in terms of the number of targets assigned to him and the number he destroys, or the number he permits to reach some minimum distance from the center. Likewise, the number of aircraft or missiles he assigns against these targets can be considered a consumption measure.

To measure the center's performance, the performances of the individual operators can be combined; for example, if each of the 6 operators had 12 targets to handle (a center total of 72) and the total targets destroyed were 63, then the center's performance would be assessed as 63/72 or 87 percent effective.

4. Example. Amount achieved is not an easy measure to apply in the TACC context because, strictly speaking, there is no product of TACC operations. It is, however, possible to apply the s/n ratio to individual functions such as air intercepts with individual aircraft (because these are essentially discrete operations). Other TACC activities (such as communicating or coordinating) cannot be evaluated objectively as successful or unsuccessful.

A consumption measure can also be applied to the number of aircraft launched by TACC to contain an aggressor force or the number of weapons it assigns to counter a threat. Such a consumption measure, however, applies only to a subset of TACC activities (countering a threat) and can be used to describe the cost of that subset alone.

5. Summary. The following steps should be followed in deriving a measure of amount achieved and/or consumed.

a. Determine what the operator's responses are.

b. Determine that they are discrete and quantitative (or capable of being scored quantitatively).

c. Determine that they are meaningfully related to the overall goal of system performance.

d. Specify a number that expresses that relationship.

Subjective Measures and Methods

Discussion

Superficially, the ideal measure is objective and quantitative. Although objective measures are necessary, they are insufficient because the very few dimensions they possess (e.g., time and amount) do not permit them to describe performance fully. Simply to know that operators have achieved a certain level of accuracy or take a given length of time to perform a task does not tell us all we wish to know. This is particularly true if operator performance exhibits deficiencies; the evaluator will wish to know what causes these and a single number or even several numbers will not usually supply this information.

In addition, the performance one wishes to measure may be difficult to specify or secure in objective terms. Suppose, for example, that one of the criteria to be applied in evaluating a new reconnaissance vehicle is comfort in riding the vehicle. It is unlikely that one could measure this quality solely by objective means.

Under these conditions the evaluator must apply subjective methods. One can in fact lay down a principle: Any evaluation will be incomplete unless subjective data are gathered in addition to those of an objective nature.

Subjective data are secured primarily through human judgment, but the demarcation between objective and subjective methods is sometimes difficult to specify. For example, if a data collector usually counts the number of holes in a target on the firing range, is this objective or subjective? The measure is secured by perception, but the act of counting is ordinarily so unequivocal that one would hesitate to apply the term subjective to it.

Objective measures must be quantitative ones; subjective measures may be quantitative or qualitative, depending upon how they are treated. Subjective measures are not automatically qualitative, since there are many ways of quantifying them (e.g., see later section on ratings).

A more serious shortcoming of subjective measures is the fact that they are more inaccurate than objective measures since the former depend more on human observation, which is often imprecise. Inconsistency (lack of reliability) is another defect; the individual who supplies subjective data may be inconsistent from one measurement to another, and of course individual observers vary among themselves. This variability may be greater or less, depending upon what is being observed. Much subjective variability results from the fuzziness of subjective criteria; for example, if the performance criterion for a squad is that it should be "highly coordinated," what is one to look for to arrive at this conclusion? If subjective criteria could be specified precisely, subjective data would be as precise as objective data.

One cannot, however, avoid using subjective methods for these reasons:

1. They provide data that cannot be supplied by more objective methods. For example, it is essential to the evaluator to determine how

the personnel whose performance is being evaluated feel about the test conditions and their own state relative to test inputs (e.g., fatigue, motivation, difficulty of the situation, problems experienced, etc.). That is because how the operator views the test situation may well determine how he performs in that situation.

2. Some data can be secured only through subjective methods. For example, it may not be clear exactly what has taken place during the operator's performance; questioning the operator can help clarify this. The factors impacting on the operator's performance may be invisible except to him.

3. Subjective methods are often less expensive to employ than objective ones. Where objective measures require expensive instrumentation which is difficult to operate and maintain (in a field environment, for example), it may be necessary to seek the same information in a different manner. There are, however, personnel costs associated with the use of subjective techniques. Subjective methods can be used to secure data on the following:

a. Description of performance (what took place).

b. Determination of performance effectiveness (by means of ratings).

c. Factors subjectively experienced by the operator that affect his performance (e.g., fatigue).

d. Adequacy of system characteristics such as the procedures used.

e. The operator's internal state (e.g., motivation).

The Interview

1. Definition. The interview is one of the most common methods of securing subjective data. In essence, it is simply asking questions of a respondent orally and noting his answers. The interview is difficult to apply well because it is the least structured subjective method (often taking form as the interview proceeds) and because it involves interpersonal relationships between the performer and the interviewer (e.g., the performer's reactions to the interviewer's manner).

Interviews may take various forms:

a. Individual (one interviewer, one performer). This is the most common form of the interview.

b. Team (one interviewer, several performers being interviewed as a group). This occurs only in connection with team operations.

c. The interview may be combined with other subjective methods such as rating scales, which are completed by the interviewee.

d. The interview may be combined with a demonstration of performance by the operator.

2. Use Factors. An interview should always be held at the completion of the test; no performance evaluation should be considered complete without it. However, the interview cannot substitute for more objective methods. It cannot, for example, be used to verify that performance satisfies system requirements or to predict future operational performance. However, it can help to diagnose problems encountered in testing and to identify required personnel resources and training.

Since the interview is directed at securing additional material concerning prior performance, the interviewer should also have observed that performance.

What types of information would one wish to secure from the interview? Literally anything about which one wishes an opinion from the performer can be secured from the interview, but the following questions are the most common. (Note: These are not necessarily phrased in the form in which they would actually be asked.)

a. What did the performer do while he was performing? For example, when the operator field stripped his machine gun, did he remove the magazine before he broke it down? Unless a record of the subject's performance was made automatically (e.g., by videotape camera), the interview may help to check the accuracy of the data collector's observations.

b. Why did the subject perform as he did? For example, why did the squad leader select one reconnaissance route rather than another? If the performer made errors, why did he?

c. What knowledge does the performer have about the principles and information that should have guided his task performance?

d. What test conditions (e.g., night/day) affected the performer most and why?

e. How well did the operator think he performed? If there are any significant discrepancies between the subject's performance as it was observed and his own evaluation of that performance, why?

f. With what tasks did the operator experience the most difficulty? Why? What factors does he feel contributed to those difficulties?

g. In a team operation, how was responsibility divided among the members?

h. Citing a specific factor of interest to the evaluator (e.g., reduced visibility, logistics, etc.), what effect did that factor have on the operator's performance?

i. Does the subject have any comments at all that he wishes to make about any aspect of the test or his performance? (This is an open-ended question, usually asked at the end of the interview, and serves as a prelude to closing the interview.)

The interview should be partially structured and partially free to vary as the subject's responses suggest. Those key questions for which the investigator thinks he needs answers should be developed in advance (although they need not be written out). A key question can be used to initiate the interview. The subject should be allowed to expand on topics of interest to him but, when the interview tends to wander, the interviewer should bring its direction back to the point by asking another of the key questions. If the subject appears reticent, the interviewer can use his prepared questions to stimulate his responses.

Where the performance being evaluated is that of a team, the interview should be conducted with the team as a group. If the team consists of many members, it may be unfeasible to interview each one individually and, in any event, individual interviews do not permit the evaluator to record team interactions.

Since performance about which the operator is being evaluated is technical in nature, it is essential that the interviewer know at least the gross technical details of the task, although he need not be as proficient as the performer in that task.

The interview is commonly held after the task has been performed or at some convenient interval in system operations. Where a series of tasks is being performed, the interview should not, if possible, be postponed until every task is completed, because then the subject's memory for events is reduced. It may not be feasible, however, to hold an interview immediately following the performance of each individual task. Obviously, one cannot break in on an on-going system operation to interview; consequently, the interviewer should look for "natural" breaks in the flow of system events to ask his questions.

The subject should be interviewed concerning each task he has not performed previously. If he repeats the same task on several occasions, and task conditions remain the same, it may not be necessary to interview repeatedly. However, it is advisable to secure two interviews (separated by an interval) for each task performed by the operator to determine the consistency of his interview responses. If test conditions do not change substantially on repeated performance of the same task, the answers the interviewer gets on the first two interviews will probably be repeated in subsequent interviews, and the performers will find the questions boring. Under these circumstances, the scope of the interview can be reduced. If there is some expectation that subject responses will change in subsequent task performance (because of learning, for example), the subject should be reinterviewed, although the length of the interview can be shortened when it is repeated.

The length of the interview should depend on how much the performer can tell the interviewer, but 20 minutes is a good average length. Beyond 30 minutes, the subject tends to become fatigued.

The conditions under which the interview is conducted are important. Ideally, it should take place away from the test operation (although not too far away lest one has to account for transportation time), in a fairly quiet

place, with no interruptions or distractions. A specific room for the interview would be most desirable since this helps to emphasize the importance of the interview to the performer. The fact that an interview will take place should be made known to the subject in advance. He should be reassured at the outset that nothing he says will be held against him, and that he will not be identified in records or reports. He may ask for feedback about his performance since he is understandably concerned about his proficiency in an evaluation. He should be told that the interviewer is not the evaluator but that it appears to the interviewer that the subject's performance was certainly adequate.

The most convenient way of conducting an interview is to tape record it but, if it is a very short one, or the interviewer is highly skilled, it can perhaps be handled by note taking. As far as possible, interview responses should be recorded verbatim without the interviewer making any attempt during the interview to analyze them.

4. Problems. One major advantage of the interview is that it is an interpersonal situation. It therefore can be used to motivate the performer to communicate by suggesting that his performance is considered important. Beyond that, many people feel less constrained in responding orally than they do in writing. Another advantage is that the flexibility of the interview permits the subject to partially control it by selecting or emphasizing topics as he wishes.

One of the disadvantages of the interview is that it is usually a one-on-one procedure, which means that a great deal of time can be consumed with this method. Since conversation is not very structured or standardized, the analysis of the material gathered is more difficult because the freedom of the procedure permits the interviewee to include irrelevant and unimportant responses.

5. Example. Interviews in the TACC evaluation situation could be either individual (one interviewer/one respondent) or group (one or more interviewers/the TACC team or some major part of it). Because TACC is a team operation, it would be reasonable to interview the team as a group. Alternately, since TACC personnel perform individual (although overlapping) functions, it would be possible to interview at least the key personnel of the team individually.

The most reasonable time for the interviews was at the end of the work shift, because it was impracticable to break into a "loose" operation like TACC that has no natural break points. The interviews, conducted by those who had been observing the previous work shift operations, were held in an administrative area located next to the TACC shelter.

It was eventually decided to hold an initial group debriefing session after the first work shift, to be followed by relatively short individual interviews with key personnel (excluding auxiliary personnel such as map and status board operators).

Interviews focussed on determining where difficulties had arisen or where problems might arise (e.g., points of heavy target inputs). The group

interview began with a recapitulation of the major events as these had occurred during the work shift. As each activity was reviewed, the individuals involved were invited to comment on the accuracy of the recapitulation and to supply additional information concerning these events. Because this was a team interview, special attention was paid to coordination within TACC and between TACC personnel and other organizations. The TACC team was invited to report any difficulties they experienced and to suggest improvements that could be made in operating procedures, technical data, etc. The interviewer took notes but the proceedings as a whole were tape recorded.

Subsequent individual interviews were held with key personnel. These dealt in greater detail with the specifics of their individual work stations and functions. Because the group interview had preceded the individual ones and had, presumably, elicited major problems, the individual interviews were substantially abbreviated.

The TACC evaluation ran daily over a period of approximately 6 weeks. During this time, a group/individual interview was held at the beginning, midway through, and at the conclusion of testing. The Chief Evaluator felt that such a schedule was not an inordinate imposition on TACC personnel and was actually the minimum he should have to gather interview data.

6. Summary. The following steps should be performed before conducting the interview:

a. Determine the information you wish to secure from the performer.

b. Develop key questions to elicit that information.

c. Within bounds, allow the interview to "go" the way the respondent wishes it to go.

d. Observe the task performance for which you are interviewing.

e. Set a maximum interview length of 30 minutes.

The Questionnaire

1. Definition. The questionnaire is a more structured form of the interview, structured because the questions are written and cannot be modified for the individual respondents.

2. Information Provided. The uses of the questionnaire are the same as those of the interview: diagnosis of problems encountered in testing and the identification of required personnel resources and training.

3. Use Factors. The questionnaire has certain advantages over the interview. Because the questions are written, the questionnaire can be administered to many individuals at one time, which makes it more economical. (Obviously, the questionnaire is particularly suitable for teams.) Since the questions asked are identical for each respondent, the variability between interviewers (which is inherent in the interview) is avoided. Because all subjects respond to identical questions, it is easier to compare responses

made by individuals. The consistency with which the individual answers can be tested by asking the same question in different forms. Although written materials (e.g., rating scales) can be administered as part of the interview procedure, it is usually easier and more convenient to administer these additional materials as part of a questionnaire.

The structure in the questionnaire also makes it somewhat easier to analyze its data. The questionnaire can be more readily tried out and its items modified to improve their precision than can the same questions in interview form.

The questionnaire has disadvantages, however. Since the questions asked are presented in a fixed order and with fixed content, the subject has less freedom in responding. If he should wish to comment on some topic not included in the questions, he cannot do so, although it is possible to pick some of this material up by providing open-ended questions asking for such comments. The subject may not understand a question and has less freedom to ask for an interpretation, although the questionnaire administrator is often available to provide guidance. The space permitted individuals in which to write is necessarily limited; many people who do not like to write will be less responsive than if interviewed. Because of the impersonality of the questionnaire format, subjects may be less motivated to respond fully. Respondents often pay less attention to and devote less care to completing a written form. The questionnaire administrator cannot ask for explanatory material when he desires it.

For all these reasons, the questionnaire is less preferred than the interview as a means of gathering data. However, it can be used to collect the same types of information as the interview (for a list of these, see the preceding material on interviewing).

4. Problems. In both the interview and the questionnaire, but more so in the latter, detailed instructions on how to respond must be provided to the subjects. Information about who has developed the questionnaire and for what reason should be provided. Questionnaire administration requires a larger office area because one is almost always dealing with a group. Writing surfaces and pencils must be provided.

Like the interview, the questionnaire is also administered after test performance. However, the events to be reported on are more likely to cover a longer period of test time, because it is unfeasible to convene a group too frequently. When questionnaires are very lengthy or where subject time is very precious, they may be administered on a "take home" basis, with the subject responding when he has time; however, this procedure is not desirable because the manner in which the respondent completes the form is uncontrolled. Because of the general reluctance to write at length, the questionnaire should be as short as possible while still soliciting the desired information.

All other provisions that apply to the interview apply also to the questionnaire.

5. Example. Because of the number of TACC personnel, which made interviewing expensive, evaluators relied heavily on the questionnaire as a source of data. On a predetermined schedule (at least once every three shifts) questionnaires (requiring, on the basis of tryout, not more than 15 minutes to complete) were handed out and personnel were asked questions such as:

a. How would you rate your performance on this shift compared to that of the preceding two shifts?

b. Were there any times during the shift when you felt you could not keep up with the input flow? (If so, describe.)

These questionnaires were reviewed and, if any response appeared to warrant further explication, interviewers contacted the respondents. The questionnaires supplemented the observational data collected during the test by allowing subject personnel to report data to which they alone had access.

6. Summary. The following steps should be followed in applying this method:

a. Use the questionnaire when it is necessary to collect data from many subjects and/or sufficient interview time is not available.

b. Reduce the length of the questionnaire as much as possible.

c. Try out the questionnaire, analyze the try-out responses, and interview respondents concerning their understanding of the questionnaire items.

d. Provide detailed instructions on how to complete the questionnaire.

e. Wherever possible, administer the questionnaire in the presence of an evaluator who can answer respondents' questions.

Observation

1. Definition. Observation is the collection of data primarily through perception and analysis of the event being observed. The emphasis here is on the adverb "primarily," because all of the methods discussed so far involve at least a minimal degree of observation of, for example, data collection instruments. Observation in the sense in which we refer to it is closely associated with ratings (to be discussed later).

2. Information Provided. The importance of observation is illustrated by the fact that it alone can supply data for all the purposes of an OST (see Table 3-1). More specifically, it can be used to:

a. Describe what events/responses occurred (what took place, who did it, when and with whom did it occur)? In highly proceduralized performance situations, this use of observation may be minimal.

b. Determine the frequency of system events or operator responses.

c. Determine the accuracy (or errors) with which these events/responses took place.

d. Infer certain qualities in performance and make judgments about that performance, as in the case of rating the efficiency of performance.

3. Use Factors. Observation is not the only method available to supply these data. Except for ratings, automated methods can be applied in each case (e.g., videotape recording) to describe what took place. Observation is merely one of the evaluator's options. Moreover, it can be used with instrumentation. For example, a videotape camera can be used as a back-up to direct observation when inputs are so numerous or so rapid that the observer cannot deal with all of them.

Observation is therefore an alternative to automatic measurement. Where instrumentation is not available or feasible (from an engineering or cost standpoint), observation becomes an essential means of gathering data. For example, when a system is evaluated during field exercises, observation may become a very desirable alternative to instrumentation.

4. Problems. Observation appears to be so "natural" a method that the test planner may tend to overlook the problems it brings. One problem referred to in previous discussions is the definition of what constitutes the system events or personnel responses to be observed. The written description of a task often varies in some dimensions (e.g., flow of inputs, team member interactions) from the task observed during performance. Written language is unable to describe fully the richness of the task as performed, especially when the task is complex (e.g., continuous or involving a team) or consists largely of covert behavior. The observer must be able to recognize (a) when the task starts and when it is completed (relatively simple), (b) which operator responses are correct and incorrect and why (somewhat more difficult), and (c) which task qualities are relevant to the purpose of the evaluation (much more difficult). For example, to determine that fatigue is affecting performance is often extremely difficult, because fatigue indices are usually not obvious.

Extensive experience in performing the task to be observed is highly desirable, but not necessary, although the observer lacking any experience must be given some training in how to perform the task. Even when he is experienced, the observer's experience of the task as he has seen it performed may not correspond entirely with the task as performed during the evaluation. For this reason, training in observation is essential.

Observation is not merely perception ("seeing"); it also involves analysis and interpretation, by the selection of relevant task characteristics. Often this analysis/interpretation is completely unconscious, but training can make the observer more aware of it. One goal of observer training is to make relevant task dimensions more visible to him.

The observer himself is the limiting factor in observational accuracy. If many task events occur rapidly or concurrently, the observer may not pick them all up. He sees, but he may not be aware of what he sees. The well known unreliability of witnesses to criminal actions is a case in point. When the observer is time or event-stressed he may:

a. Select a subset of task events to report, ignoring the remainder. What he reports may therefore be incomplete and only partially relevant.

b. Use gross rather than detailed reporting categories.

These difficulties are particularly evident when the observer rates the operator's performance. Ratings are apt to be based on several dimensions (e.g., the operator's speed, accuracy, coordination, etc.). If the observer rates a task performance as "adequate," what does this mean? Gross categories of performance, such as adequate, acceptable, satisfactory, etc., usually involve several elements, some of which the observer may ignore in making his judgment. Did the observer actually note these elements and did he weight them correctly when he rated the performance?

As the complexity of the task being observed increases, it becomes more desirable to have two observers whose results can be correlated for consistency. However, this increases the number of required data collectors.

Observers must, therefore, be familiar with the tasks to be evaluated and they must have an opportunity to practice observing and reporting their observations. Observations made during training should correspond as closely as possible to the conditions under which actual observations will be made. If possible, criteria of observational accuracy should be established and used during training to evaluate the observer.

Training also presupposes that instructional materials have been provided the observer. These should include the following:

a. Specification of the task characteristics the observer should be looking for, together with operational definitions of these characteristics.

b. A properly designed form with which to report the events observed.

How the observer reports his observations may significantly affect his accuracy. The following are possible observational formats (not all of them are equally desirable):

a. The observer may attempt to write down or tape record everything he has observed or heard. This is undesirable because he will probably leave out certain events.

b. If the task is highly proceduralized, he may use an operating procedure as a guide to task performance and check off each step as the operator

makes it (see, for example, Figure 3-3). However, this may be difficult if the task is continuous or the sequence of operator responses is variable (contingent).

c. He may develop a checklist of significant aspects of the task (those actions most likely to be performed erroneously or those that may be most difficult for the operator) and use this as a guide to what he should observe.

d. He may combine formats b and c, using an operating procedure as a general guide and noting all errors or other behavior of interest as these occur.

The format selected will depend at least in part on the nature of the task to be observed.

If the observer is physically to see enough, he must be close to the task being performed. If he is too close, however, he may intrude upon the performance and thus distort the way in which the task is performed. Every observer must be considered as potentially obtrusive, even when the operator knows why he is being observed and is familiar with the observer.

5. Example. In the TACC evaluation, observations were made as part of the process of rating individuals and the total team on their efficiency in monitoring airspace, launching interceptors, coordinating assault forces, and interacting with other team members and other air defense organizations. Only the 11 key personnel were observed and observations were made only on a sampling basis. Prior to the start of the evaluation, test planners met at intervals over several weeks to define as precisely as possible what was to be observed. Questions raised at these meetings were:

a. Is the desired behavior reliably observable?

b. Is it meaningful in terms of the evaluation purposes?

c. What forms will be used to record the data?

d. How will the observation/ratings be analyzed?

Rating forms to be discussed subsequently were developed to include these definitions. The first three or four observations were considered as essentially practice sessions and the results achieved with these were analyzed by the test planners for consistency, difficulty in making the observation, and accuracy of the results. Consequently, the observational methods with which the study began were somewhat revised before planners were satisfied that valid data could be secured.

6. Summary. In making use of the observational method, the following steps should be followed:

a. Define as precisely as possible what is to be observed.

b. Select subject matter experts as observers (because these are easier to train, having had prior experience) and provide systematic observer training.

c. Develop a reporting form that provides cues as to what is to be seen and that indicates as precisely as possible the categories with which the observation is to be reported.

d. Measure the consistency of individual observers on repeated observations and the consistency between observers. Reject those observers who are not consistent in their performance.

Ratings

1. Definition. Ratings are the means by which observational judgments are quantified. There are several kinds: (a) ratings by a data collector/evaluator, (b) supervisor ratings (where a team is involved, one of whose members is in charge of team activity), (c) peer ratings or ratings by one member of the performance of other team members with whom he has interacted, and (d) self report ratings in which the operator judges his own performance or his own feelings regarding an event or performance that he has accomplished.

The difference among these ratings is the rater's involvement in the activity being rated. In self ratings, he evaluates his own performance; in peer and supervisor ratings, he is only a partial participant in the activity being evaluated, and in data collection/evaluation ratings, he is a nonparticipant. Most ratings in the OST context are data collector ratings, but there is also value in the other ratings. Conceivably, the more involved the rater is, the more valid his ratings, but one must also consider bias.

Ratings are quite popular in personnel performance measurement for two reasons: (a) assigning a number to a subjective assessment makes it possible to treat that assessment quantitatively, and (b) most rating scales used for operational evaluation can be developed quickly, are easily used, and require little or no specialized training.

Having explained rating popularity in this way, it is necessary to say also that applying numbers to a subjective assessment does not make that measurement more accurate or valid. To utilize ratings correctly does, in fact, require extensive development, validation, and training of the rater.

This discussion does not pretend to be a highly technical analysis of ratings and rating scales, about which much theoretical material has been written; only the essential points relevant to OST are covered here. The reader interested in more detail should see Guilford (1973).[2]

---

[2]Guilford, J. P. Psychometric Methods. New York: McGraw Hill, 1973.

2. Information Provided. When a judgment is made that a particular performance is satisfactory or unsatisfactory, ratings can be used to verify performance. To the extent that verification represents how the system will do operationally, ratings can be used to predict operational performance. The same data can be applied to future systems. When a performance rating reflects on the training required to achieve that performance rating, it can also be used to provide data for future training.

Literally anything can be rated, assuming one can scale it. Ratings should be used, however, only when an objective method cannot be found, unless cost or feasibility considerations make the latter undesirable. Ratings are most often employed for the following purposes:

a. To evaluate how well someone is performing a task (this rating can be made by an observer or by the performer himself).

b. To evaluate some quality of task performance, (e.g., coordination of squad members).

c. To evaluate the adequacy of some feature of the system being evaluated (e.g., displays, procedures, and logistics).

d. To evaluate the effect of some condition (e.g., visibility) which impacts upon performance (e.g., detection of targets).

e. To evaluate the output of task performance (e.g., the adequacy of a tactical decision).

3. Use Factors. Since ratings depend on observation, rating validity (the extent to which the rating mirrors real phenomena) depends on the observer as a measuring instrument. Everything said in the previous section on observation therefore applies to ratings.

There are several types of rating scales, but since the graphic scale is the most popular and the most adaptable to OST, this is the one that will be described.

The graphic rating scale can be designed in different ways by taking a straight line (representing the scale along which the variable varies) and combining it with descriptors that represent the values of the variable being scaled. The line can be segmented in units or it can be continuous. If segmented, the number of segments can be varied. The line can be placed horizontally or vertically. The number of scale points can be varied.

A number of guidelines for development of the scale have been laid down:

a. The line forming the scale should be 5-6 inches long.

b. Continuous rather than segmented lines are preferred because the former suggests that the variable being measured is continuous.

c. Where several scales are used to measure individual dimensions of the variable, the ends of the scale representing the "good" or "high" aspects should all be in the same direction.

d. If the scale is vertical, put the good or high end of the scale at the top of the page.

e. The descriptors of the various points along the scale need not be equally spaced along the line, although this is general practice, because usually one does not know how the variable being rated deviates from linearity.

f. The descriptors are placed opposite the scale points which they describe.

g. Descriptors whose meanings or values are so extreme (e.g., never or always) that they will almost never be applied by the rater should not be used.

h. When the variable being evaluated is bipolar (i.e., can be described either positively or negatively, as, for example, speed ranges from slow to fast), the neutral point is midway along the line.

i. The number of scale points generally vary from 5 to 9. Fewer than these leads to grossness and lack of precision in the rating; more than these cannot be meaningfully discriminated by the rater. The precise number to be used depends on the nature of the variable and the evaluator's ability to think of distinctly different steps along the scale for that variable. A 5-point rating scale is most common, however.

Graphic rating scales are fairly simple to develop and easily administered. This says nothing at all for their validity or the process of validating them, which can involve considerable effort and which is why most scales used in OST are not validated. If used for self rating performance, they are interesting to the rater. They are easily completed, since the rater does not have to perform any mental calculations. They provide for as fine a discrimination as the rater is capable of and as fine a degree of scoring as the evaluator desires.

The first (and most difficult) steps in the development of the rating scale are (a) to conceptualize the performance variable (e.g., speed or degree of comfort experienced) to be rated in terms of a continuum, and (b) to segment that continuum into psychologically equal intervals.

The performance attribute to be rated must be observable by the observer or the task performer, which is simply another way of saying that whatever is to be rated should be capable of being rated. If scales are developed for an attribute which the rater cannot validly observe, the resultant ratings will be invalid, even though numbers have been applied.

Everything said in Section Two concerning criteria and measures in the Personnel Performance Test Plan applies also to the development of ratings. Before the rating scale is developed, it is necessary to operationally define the performance variable to be rated. It is easy to scale any performance in terms of values such as excellent, very good, good, adequate, slightly inadequate, poor, and terrible (a seven-category scale), but what do these phrases mean? Suppose one wished to rate ability to perform ground reconnaissance, using the previous categories. What would a rating of adequate in this situation depend on? Speed of detecting certain terrain or man-made features? Speed of reporting information? The amount of detailed information provided in the report? And if so, how much speed or information denotes adequacy? Obviously we come immediately up against the problem of criteria. Where objective criteria are not available, they can be developed using the techniques described in Section Two. Because the criterion must be scaled to make it suitable for rating, it is necessary to fractionate it into equivalent parts, which means that a criterion must be developed for excellent, very good, good, etc.

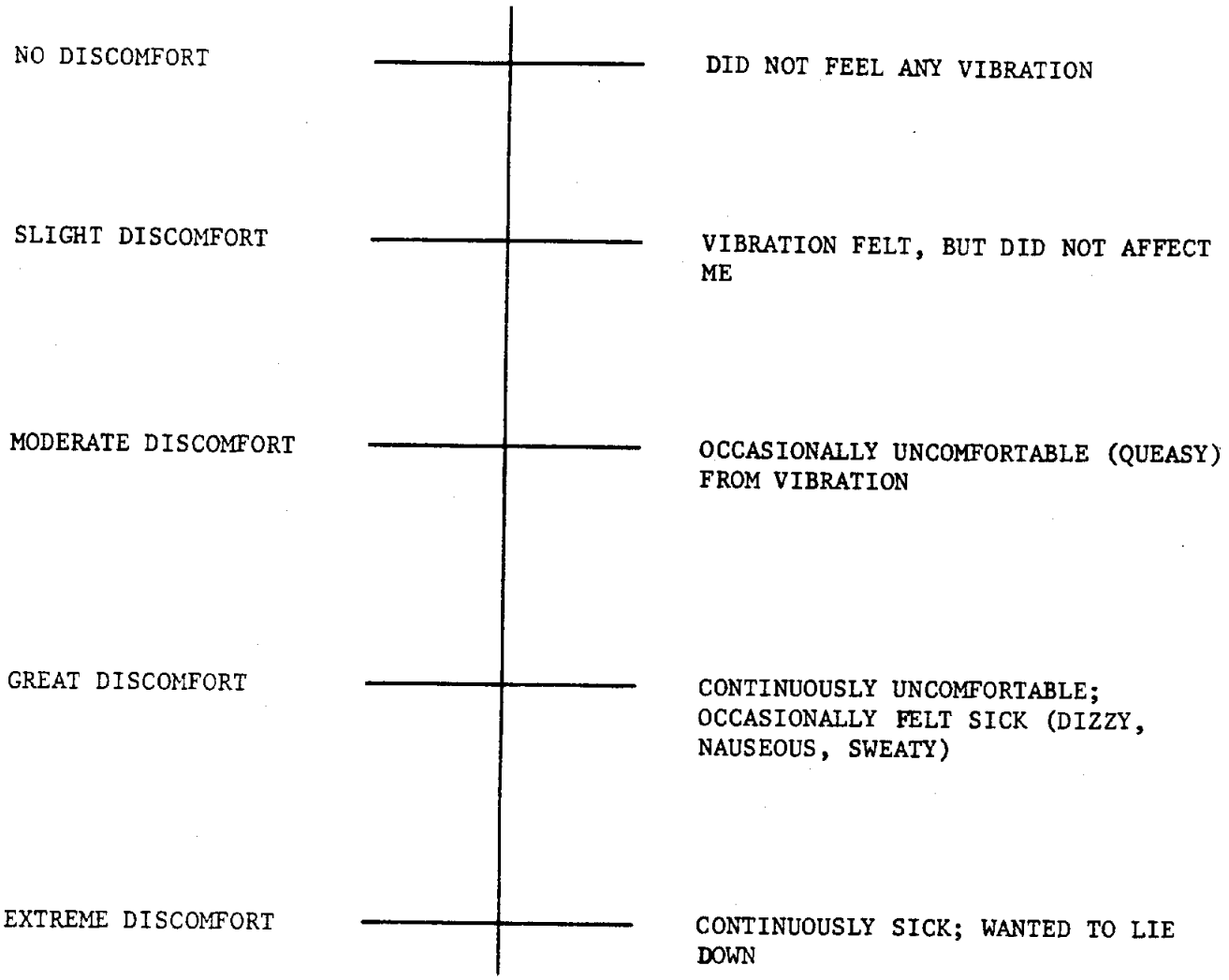
The rating is particularly useful when performance depends on multiple correlated dimensions (such as adequacy of ground reconnaissance). It may not be feasible to develop individual, independent measures of each dimension and the integrated criterion may not be susceptible to instrumented measurement. In this case a global rating is appropriate provided the rater is aware of the dimensions making up this rating.

As an illustration of the preceding concepts, assume that the variable being rated is the amount of vibration experienced by troops being transported in a LVTP-7. Note that this is not the same as the degree of vibration physically occurring in the carrier; the latter can and should be more effectively measured by means of accelerometers. What is being rated in the sample scale (Figure 3-4) is the individual Marine's experience of that vibration (this is a self-report scale), which cannot be physically measured. This is a good example of a subjective measuring tool which complements an objective one. Both objective and subjective measures are desirable, since the experienced vibration and its effects may not be linearly related to physical vibrations.

The first thing to do is to decide on the dimensions of the vibrations (i.e., the basis for making judgments) as conceptualized in Figure 3-4. There are two dimensions in the Marine reaction to vibration: (a) the frequency of physical symptoms resulting from the vibration (e.g., never, occasional, continuous), and (b) the severity of those symptoms, (e.g, discomfort, sickness). These dimensions can be varied interactively on a scale of discomfort. (Note that, since vibration can have only negative effects, this is not a bipolar scale, so one need not worry about the neutral point between the two poles.)

Once these dimensions are conceptualized, the extremes of the scale values are developed: no discomfort and extreme discomfort. Intermediate values can then easily be supplied: slight, moderate, and great discomfort. This gives us a five-point scale.

Since discomfort is represented by more than one dimension, it is necessary to supply more detailed descriptors at each scale point, as shown in Figure 3-4.



Definition: Vibration discomfort is any feeling of unpleasant physical sensation experienced by personnel as a direct result of vehicle vibration.

Figure 3-4. Vibration discomfort rating scale.

Note that the two dimensions (frequency, extent of symptoms) interact, as in the scale item "great discomfort" where the Marine reports that he was continuously uncomfortable and occasionally sick. The respondent can select any point on the scale (including those between descriptors) that represents his feeling of discomfort.

A number of questions can be asked about such a rating scale. Do the two dimensions (frequency and severity of symptoms) truly represent the elements of discomfort as actually experienced by personnel? This can, perhaps, be determined in advance by questioning personnel who have experienced discomfort in transport vehicles.

A second question is whether the intervals between scale values are equal (which is the way the sample scale was devised). Is the difference between NO DISCOMFORT and SLIGHT DISCOMFORT the same as the difference between GREAT and EXTREME DISCOMFORT? This is a tough question to answer. If the range of discomfort is equally distributed in personnel in real life, one should find an approximately equal distribution of scale responses if one tested a large enough sample of respondents. However, in actual practice, one simply assumes equal intervals and lets it go at that.

The scale points, it will be noted, have not been numbered (to avoid biasing the subject), although this is entirely possible. In this case NO DISCOMFORT represents 0 (zero), EXTREME DISCOMFORT, 4. Since the respondent can mark anywhere along the scale, one need only measure the distance of the respondent's check mark from the nearest descriptor to secure a quantitative value for his response. For example, if he checks half way between SLIGHT (1) and MODERATE DISCOMFORT (2), he would have a score of 1.5 on this scale. Precision of scale measurement can be anything the evaluator desires, but in a relatively crude instrument like this one, he would probably not attempt to measure to more than tenths (e.g., 1.5, 1.6, 1.7, etc.). Precision is not the same as accuracy or validity.

An attribute to be rated should ordinarily be introduced with a name and a definition phrased in operational terms; these should provide cues to the rater to permit him to recognize the attribute. The attribute name without accompanying definition is likely to be misleading.

Rating scale descriptors such as "occasionally uncomfortable (queasy) from vibration" (see Figure 3-4) have the double function of supplementing the definition of the variable being rated and of providing anchors or milestones to guide the rater in making quantitative judgments. The following criteria should be followed in constructing scales:

a. Clarity. Use short, simple, unambiguous terms.

b. Relevance. The descriptor should obviously be consistent with the attribute being rated and its definition.

c. Precision. A good descriptor applies to a very short range on the variable continuum. There should be no doubt about its rank position relative to other descriptors.

d. Variety. Vary the language for descriptors at different scale points.

e. Objectivity. Descriptors that imply that a certain behavior is good or bad, desirable or undesirable, should be avoided.

f. Uniqueness. Descriptors should be unique to the attribute being rated. Where possible, avoid very general descriptors such as "excellent" or "adequate."

Experience has taught us a good deal about raters and their peculiarities:

a. Individuals differ in their capability to rate others.

b. Two ratings by the same rater are no better than one because the rater tends to repeat his errors a second time.

c. Raters need sufficient time to make their judgments.

d. Raters do better if they know the task whose performance they are evaluating.

e. Self-raters tend to overestimate more than underestimate their performance on the variables they are rating.

f. The assurance of the rater is important. Judgments of which he is very sure are much more reliable than other ratings.

g. Different raters often use different criteria in judging. This makes systematic training in rating all-important.

When finally developed, the rating scale should be tried out to see how well it performs. Although complex experimental and statistical procedures for scale validation exist, these are usually not feasible in an operational test situation (e.g., lack of time). Where possible, however, the prospective scale should be applied to two separate subject samples, known in advance to vary on the performance variable being rated, to see if the scale actually differentiates between them. For example, to rate the adequacy of maintenance on an equipment rating scale, evaluations of experienced technicians and students in training might be compared to see if the scale will provide differential answers (as it should, because one would logically expect these two types of personnel to behave differently in troubleshooting).

If this is not possible, the scale should at least be tried out with a number of raters to ensure that the meanings of terms are understandable, that there is no confusion between descriptors, and that, in general, problems will not be experienced in using the scale.

Since the measuring instrument in the rating scale is the rater himself, it is critical that he be trained in the use of the scale before he actually applies it. Training should involve the following:

a. Analysis of the operational definitions of the performance variable and the individual scale points.

b. Examination of the criteria inherent in the scale.

c. Practice in rating the performances to be evaluated.

d. Comparison of rater consistency with the results of others who are independently rating the performance. We deliberately have not mentioned rater accuracy, because this assumes an independent measure (which one assumes to be valid) of the variable being rated. Such independent measures are rarely available.

e. Training in the avoidance of common errors made in using rating scales.

4. Problems. Experience with ratings over the years has identified a number of errors commonly made by raters that lead to rating inaccuracies.

a. The error of leniency. Raters tend to rate those whom they know well or with whom they are ego-involved higher than they should. Some raters are "hard," others are "easy." Easy raters are more common. Any tendency to be unduly easy or hard in rating should be determined during training practice sessions and this tendency should be brought to the rater's attention.

b. The error of central tendency. Raters hesitate to make extreme judgments and thus tend to "bunch" their ratings in the center of the scale. This reduces the discriminating power of the scale.

c. The halo effect. Where several dimensions of the operator's performance are rated separately (e.g., speed, accuracy, coordination), the rating given any single dimension is likely to be affected by the rater's general impression of the individual being rated. Halo effects are most likely to be found in performance attributes that are not clearly (objectively) defined and that are not easily observed.

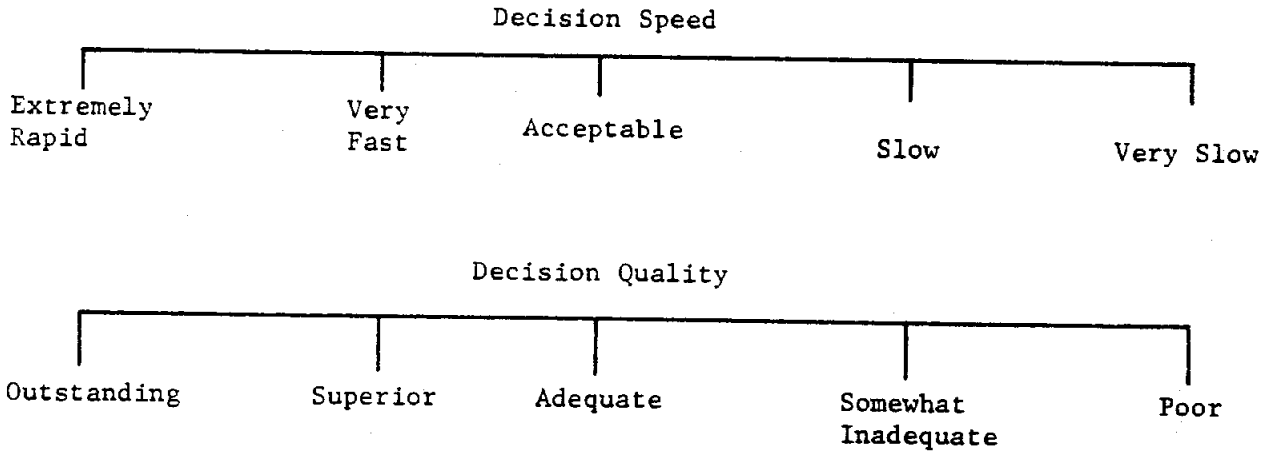
5. Example. Because many of the functions performed in the TACC do not lend themselves to objective evaluation, it was necessary to rely heavily on ratings of performance effectiveness. Theoretically, these would be based on observable criteria for which a judgment could be made. A number of rating scales were developed but only one will be discussed in any detail.

Since the TACC is divided into functional areas that operate semi-independently, an efficiency rating scale had to be developed for each area. In this case, it was the Air Defense Coordinator (ADC) together with his

assistant. The first step in the development of their rating scales was the specification of the functions to be evaluated and the effectiveness criteria for these functions. The ADC functions include monitoring, supervising and coordinating control of aircraft and missiles by subordinate agencies, initiating orders to these agencies, monitoring equipment and operational status, controlling the tactical alert net, etc. Speed and accuracy are obvious effectiveness criteria for these functions, but the problem was that objective values of these criteria did not exist. Attempts to develop standard values for these criteria were ultimately abandoned because agreement on them could not be reached.

It was decided that the one critical element in the ADC's activity that cut across all his functions was his decision making and this involved two aspects: speed and decision quality. This made two scales necessary. Effectiveness was to be determined by the judgment of highly skilled, experienced personnel who had operated an earlier version of the TACC.

The effectiveness rating scales for the ADC as finally developed took the following form:



Because of the lack of objective standards for decision speed and quality, the responsibility for valid measures depended completely on the rater. For this reason, it was decided that any evaluation score less than adequate (midpoint of the scale) would be documented by the rater with specific instances of the behavior on which the evaluation was based. In that sense, the rating was supplemented by critical incident data (see subsequent discussion). These data were evaluated to justify a less than desirable score.

The rater was also required to practice making observations and his first three or four ratings were discarded as essentially practice evaluation. Ideally one would have wished to have two raters simultaneously and independently evaluating ADC performance, but the lack of qualified observers prevented this.

6. Summary. In making use of the rating method the following steps should be implemented:

a. Determine the variables for which ratings will be sought and the continuum on which those variables can be rated.

b. Develop objective criteria to define the range of values along which the variables will be rated.

c. Utilize the rating scale development principles described in this section.

d. Try the rating scale out to see how well it performs.

e. Provide systematic training to the rater in observation of the variables being rated and use of the scale.

The Checklist

1. Definition. The checklist is a series of statements that describe characteristics of (a) operator performance, (b) the equipment configuration, and/or (c) system operations. The checklist statement serves as a standard against which the evaluator makes a judgment that what is being evaluated either has or does not have the characteristic described in the statement.

2. Information Provided. The checklist is used to verify that the system satisfies its requirements. It is used most frequently to evaluate relatively static attributes of the system, such as characteristics of the man-machine interface; one finds it most commonly in the form of the human engineering checklist. The checklist is not often used to describe performance because performance involves continuous variables and the checklist statements are binary. In general, human engineering evaluations of equipment should be performed during developmental testing (prior to OST); however, the occasion may arise to use the checklist method during OST.

3. Use Factors. Like the other methods described previously, the development of a checklist presumes a criterion: that the object or performance being evaluated should possess certain desired characteristics. It is impossible to develop a checklist of qualities before having decided what those qualities should be. The checklist also presumes a value judgment: if the object or performance exhibits characteristics X, Y, and Z, it is adequate or effective in performance; if it does not, it is inadequate and ineffective.

If one were to develop a checklist to evaluate performance, it might contain statements such as the following:

a. Consults TMs frequently during checkout.

b. Strips machine gun quickly and correctly.

These statements suggest that consulting TMs frequently and stripping a machine gun quickly and efficiently are desirable qualities of task performance. What do frequently, quickly, and efficiently mean? Since these adverbs

reflect an underlying quantitative continuum (e.g., TMs might never or always be consulted), they can be measured more directly by means of time and error measures and frequency of occurrence.

On the other hand, if the evaluator uses a checklist because he cannot precisely define what he means by frequently, quickly, and efficiently, then any evaluation he might make using these terms would be highly invalid and inconsistent.

Because dynamic performance can be measured by objective methods, the checklist is not a very efficient means of performance measurement. Where, however, the object of the evaluation is static and where the qualities to be measured are nonquantitative, the checklist can be useful, provided that these qualities as expressed in the checklist are meaningful in terms of some standard or criterion. Moreover, the checklist statement does not indicate how much of that characteristic the object or performance being evaluated possesses. Hence, the checklist is purely a binary device. Because of this quality it may be misleading, since few characteristics exist totally or not at all.

The checklist is most frequently used for evaluating the human engineering characteristics of equipment. Criteria for specifying the desirable characteristics of the man-machine interface can be found in MIL STD 1472B (DoD, 1974)[3] which governs the human engineering design of all military systems. By determining the particular characteristics that represent the man-machine interface under evaluation, one can refer to 1472B and find the features to be described by checklist statements. For example, on page 59 of 1472B one finds the paragraph 5.4.1.2.1, Consistency of Movement, which can be applied to the equipment if it has controls:

> Controls shall be selected so that the direction of movements of the control will be consistent with the related movement of an associated display, equipment component, or vehicle. In general, movement of a control forward, clockwise, to the right, or up or pressing or squeezing a control shall turn the equipment or component on, cause the quantity to increase, or cause the equipment or component to move forward, clockwise, to the right, or up . . . .

The checklist items that can be developed from this paragraph are:

a. The direction of control movements is consistent with the related movement of their associated displays.

b. Movement of a control forward, clockwise, to the right, or up causes the quantity to increase, or causes the equipment/component to move forward, clockwise, to the right, or up.

---

[3]Department of Defense, MIL STD 1472B. Human Engineering Design Criteria for Military Systems, Equipment, and Facilities. 31 December 1974.

4. Problems. Using the checklist requires that its statements be interpreted. Because each statement implies a binary situation (the equipment has or does not have the specified characteristic), intermediate situations present problems. For example, the previous checklist items assume that all controls possess the direction of movement attribute. Suppose, however, that of four controls, three possess the quality and one does not. Should the item be checked off as OK or not OK?

The specificity of the checklist item also determines the precision with which the checklist evaluation can be performed. For example, page 17 of MIL STD 1472B (paragraph 4.6, Simplicity of Design) specifies:

> The equipment shall represent the simplest design consistent with functional requirements and expected service conditions. It shall be capable of operation, maintenance, and repair in its operational environment by personnel with a minimum of training.

Assuming one translates this requirement into a checklist item, it might read as follows: "Equipment design is simple, easy to operate and maintain."

In making this checklist evaluation, the evaluator must ask himself, what does "simple, easy to operate and maintain" mean? Obviously the grossness of these concepts makes it impossible to derive any meaningful results from their evaluation.

The checklist also suffers from another difficulty: its judgments cannot be handled quantitatively. Since the product of the checklist evaluation is merely a "Check" that a specified quality does or does not exist, the most one can do is to sum the number of such checks. Obviously, the greater the number of these, the better (or worse) the system or equipment is; but how much better or worse cannot be determined. The sum of a checklist evaluation cannot, moreover, be added to any other statistic describing performance. It cannot be treated like a rating because it is noncontinuous.

Checklist data are therefore merely descriptive. The checklist does, however, serve to remind the evaluator that certain qualities should be looked for. Once those qualities have been judged, the checklist evaluation pinpoints the characteristics that need improvement.

There is no standard human engineering checklist as such, because the nature of the system attributes to be examined will vary with the system. Because of the large number of qualities which any moderately sized system has, it is possible to develop a highly detailed checklist with hundreds of items; on the other hand, by describing the system more broadly, one can have a very short checklist. The length of the checklist will depend on the number of qualities the evaluator decides are critical to effective system operation; there are no general rules that can be applied.

Many evaluators (in industry) make up their own human engineering checklists, but a number are already available which can be very useful. Notable among these is HEDGE (U.S. Army, 1974),[4] which was specifically developed for the U.S. Army and therefore deals with characteristics of equipment similar to that of the Marine Corps.

5. Example. Human engineering checklists are, of course, applicable to the TACC, principally in terms of the control/display configuration of the situation display consoles, and somewhat less so to the WASDG and the manual status boards. Of these, the most important are the consoles. Among the console characteristics that might be evaluated using the human engineering checklist are the size, legibility, and intensity of symbols on the situation display CRT, the specific symbols and codes provided, and the arrangement of controls for processing information. Adequacy of procedures for performing console operations would also be evaluated using the checklist.

During the developmental testing of the TACC, the individual TACC equipments had been evaluated by means of human engineering checklists (as well as by other means). However, TACC evaluators retained the option of reevaluating these equipments if other evaluation results suggested difficulties in operating the equipment.

6. Summary. In making use of a human engineering checklist the following steps are desirable:

a. Check to determine whether the system being evaluated has previously been analyzed using a human engineering checklist.

b. If a human engineering checklist evaluation is desirable, refer to one of the already developed checklists like HEDGE to see whether it or any of its items adequately describe the system to be evaluated.

c. If already existent checklists are not satisfactory, refer to MIL STD 1472B to determine which sections of this standard apply to the system under evaluation.

d. Select applicable provisions of MIL STD 1472B and translate these statements into checklist form.

e. Prior to performing the human engineering evaluation, select qualified human factors engineers or operational personnel with evaluation experience as the checklist evaluators. Because of the subjectivity involved in such evaluations, it is desirable to have more than one evaluator apply the checklist.

f. Have the checklist evaluator study the system to be evaluated before applying the checklist.

---

[4]U.S. Army. HEDGE (Human Factors Engineering Data Guide for Evaluation). U.S. Army Test and Evaluation Command, Aberdeen, MD, March 1974.

Critical Incidents

1. Definition. A critical incident is an unusual event that occurs during system operation and that illustrates some outstanding (either positive or, more likely, negative) characteristics of the system.

Although it can describe any unprogrammed occurrence (such as an equipment malfunction), the term is generally taken to mean some task-related action performed (or not performed) by the operator. One such instance was recorded by the author many years ago during the testing of an early guided missile which was launched on a test run at a range near the city of Albuquerque, New Mexico. Standing at one of the consoles in the blockhouse over a very prominent red destruct button was the Safety Officer, whose task was to destroy the missile by pressing the button if the missile drifted off-course. Shortly after the flight began, the missile began to drift off-range. The Safety Officer was poised over his console as it became apparent that the vehicle, if it continued its flight, would impact somewhere in Albuquerque. The drift became marked, but the Safety Officer, mesmerized and sweating, did not respond. Just as the missile was about to descend on the city, someone thrust the Safety Officer aside, pressed the destruct button, and destroyed the vehicle.

The example suggests certain characteristics of the critical incident. The action taken (or not taken) is comparatively rare. Because it is unprogrammed, it is difficult to define in advance of its occurrence. It represents either extremely effective or extremely ineffective behavior (more usually the latter, because it is easier to recognize deficiencies), but it is obviously in a class distinct from more characteristic errors. Since it is unprogrammed, it is difficult to set up instrumentation to secure such data; hence, it is generally recorded through human observation, all the more so because it requires interpretation to recognize that the phenomenon is occurring.

2. Information Provided. Since the critical incident is unusual, it may imply an underlying system problem. Since it directly reflects the operator's performance, which includes his training, it may suggest certain changes needed in future training or other personnel requirements.

3. Use Factors. The incident is reported verbatim, almost like a story: who, what, when, where, why (if known), and the consequences (if any) of the incident. Since the event is relatively rare and disproportionate to preceeding events, its interpretation is by no means obvious. Consequently, the cause of the incident should be investigated subsequently by interviewing the performer responsible for the incident. Events occurring previous to the incident can be examined for clues to its meaning.

4. Problem. The critical incident cannot be quantified except by counting the frequency of its occurrence and even this does not tell us much. However, the incident can be highly diagnostic of problem areas deserving further investigation.

5. Example. As was discussed in the section on ratings, the evaluator of the TACC ADC was required to justify any less-than-adequate rating given to the ADC's performance; the justification was a critical incident that demonstrated the inadequacy. As it turned out, only a few such critical incidents were reported (primarily in the early stages of the OST). However, these were important because they occurred as a result of unusual events for which procedures had not yet been developed and hence suggested the development of those procedures.

Summary

1. Whenever possible, use both individual and team measures concurrently in order to determine the contribution of the individual operator to system output.

2. Objective measures are preferable to subjective ones, but both should be used. Subjective methods supply information that helps to explain the meaning of objective data. Any evaluation will be incomplete unless subjective data are gathered in addition to objective data.

3. After selecting a general class of measure (e.g., accuracy), the test planner must develop that measure in terms of the particular task and system he is evaluating.

4. In selecting a measure and measurement method, certain criteria should be applied. The measure should (a) be objective, quantitative, unobtrusive, and easy to collect, (b) require, where possible, no specialized psychological data collection techniques or instrumentation, and (c) cost little or nothing.

5. Reaction time is the time between the occurrence of an event and the start of the action required by the event. If a time requirement does not exist, RT is unlikely to be of value in evaluating the system unless the system is required to respond as quickly as possible.

6. Task duration is important only when a maximum duration is specified for a task or group of tasks. However, this type of measure is easy to collect.

7. All errors are not the same, differing in terms of the criticality of their effects on the task and the system output. Hence, errors should be weighted on a criticality scale. Before collecting error data, an analysis should be performed on the kinds of error likely to occur and their potential effects. A criterion of what constitutes acceptable accuracy should also precede any error measurement. Trivial errors need not be collected.

8. The frequency with which events (e.g., errors) occur should also be recorded because the data may have diagnostic value.

9. The percentage of tasks successfully accomplished by the operator (s/n ratio) is a useful measure of amount achieved by the system, provided the tasks being evaluated are discrete and are performed several times.

10. No performance evaluation should be considered complete without an interview with the performer. Although there are substantial advantages to the interview method (flexibility, interpersonal relationships), it is very costly in terms of subject time.

11. Along with the interview, the questionnaire is the most commonly used subjective method. It covers essentially the same ground as the interview but is less costly, being administered to groups rather than individuals. However, it is less flexible than the interview and more subject to misinterpretation by the respondent.

12. Observation is the means by which all data are collected; however, it is especially important in applying subjective measures, particularly ratings. Observation is extremely difficult to apply correctly unless what is to be observed is defined precisely and training is given the observer.

13. Graphic rating scales are particularly popular in subjective evaluation, because they allow numbers to be assigned to observations. However, rating accuracy depends upon the precision of the observer who is rating, and this in turn depends upon the precision of the rating criterion. The most difficult step in the development of the rating scale is to conceptualize the performance variable being rated in terms of a continuum. Training in using rating scales is essential.

14. Checklists are based on the evaluator's judgment of characteristics of operator performance, equipment or system operation. The judgment is that what is being evaluated either has or does not have the checklist characteristics. The checklist is therefore limited because of its binary nature.

SECTION FOUR--HUMAN ENGINEERING CHECKLIST PROCEDURES

This section describes a human engineering checklist which can be used to evaluate the adequacy of new equipment from the standpoint of operability and maintainability. The checklist in this section was developed to satisfy the following criteria:

1. It should be reasonably short and yet cover the essential features.

2. It should not require either specialist background or instrumentation.

3. Checklist items should be capable of being evaluated by inspecting the physical features of the equipment being evaluated or by simply measuring or observing the equipment while it is being operated.

Introduction

The purpose of this section is to present a human engineering checklist that will enable the evaluator to ensure that equipment features critical to personnel operation and maintenance of equipment have either been included or, in the case of negative features, avoided.

The checklist has been designed for use by operational personnel who are not human engineering specialists. It is based on an earlier checklist developed for the U.S. Army--the Human Factors Engineering Data Guide for Evaluation (HEDGE),[1] which was in turn based on MIL-STD 1472B.[2] The present checklist has been, however, extensively modified for Marine Corps use.

The checklist described in this Section has been based on the following assumptions:

1. If the checklist user is a military man, he should not be asked to employ criteria that demand a specialist background. Consequently, items that require a specialist background have been eliminated.

2. Criteria that are so general that one has difficulty using them for evaluation (e.g., the requirement that the equipment should be designed simply) should be avoided.

3. Relatively few equipment features can be meaningfully observed in making a human engineering evaluation. Long lists of checklist items that follow the provisions of MIL-STD 1472B step-by-step are merely deceptive; even experienced human engineers cannot properly evaluate all these features. Moreover, for many of these checklist items, even when the man-machine system is found to deviate from these criteria, the significance of the deviations (in terms of whether or not they will seriously affect operator/maintainer performance) is suspect.

---

[1]U.S. Army, Human Factors Engineering Data Guide for Evaluation (HEDGE) U.S. Army Test and Evaluation Command, March 1974.

[2]Department of Defense. MIL-STD 1472B, Military Standard, Human Engineering Design Criteria for Military Systems, Equipment and Facilities. Washington, D.C., 31 December 1974.

4. Most of the criteria specified in MIL-STD 1472B (which governs the human engineering requirements of all Department of Defense equipment) are designed as guidance to the engineer and not as evaluation criteria.

The checklist has also been shortened as much as possible to make it more acceptable to users.

Like most checklists, this one is not designed to produce a number that summarizes the "goodness" or "badness" (from a human engineering standpoint) of an equipment. Rather, it is designed to allow the evaluator to pinpoint the equipment features that require changes either in hardware or procedures. In each evaluation, the evaluator must use common sense to decide whether a particular equipment feature deviates sufficiently from the checklist criteria to significantly affect operator/maintainer performance.

In using the checklist, the evaluator must, of course, know how the equipment functions. If he does not, he should start by analyzing the operating (or maintenance) procedure for the equipment and then inspect the equipment to see if it agrees with checklist criteria. For example, suppose the checklist item is "Controls should be located under or to the right of their associated displays." In this case, if the evaluator does not know which controls are associated with which displays, he should go to the operating procedure for the answer. In most cases, if the evaluator is at all familiar with the equipment and its operation, even this preliminary step (analyzing the operating/maintenance procedure) will not be required.

Human Engineering Checklist

Obviously not every checklist item will be pertinent to a particular equipment and the evaluator will have to examine the list of items to see which are relevant to his particular equipment. Checklist criteria in this section are broken out in terms of the following categories:

1. Controls.
2. Displays.
3. Labels.
4. Workspace.
5. Stairs and ladders.
6. Handles, handholds, and railings.
7. Doors, hatches, and entryways.
8. The working environment.
9. Lines and cables.
10. Fasteners and connectors.
11. Cases and covers.
12. Access openings.
13. Maintainability.
14. Communications.

1. Controls should be located under or to the right of their associated displays.

*2. Control movement should be consistent with the related movement of an associated display, equipment component, or the vehicle as a whole. For example, a control movement forward, clockwise, to the right or up, or depressing the control, should move the equipment forward, etc.

*3. Rotary valve controls should open counterclockwise.

4. Controls should be grouped by common functions or arranged in the operating sequence (if possible).

5. Critical, frequently used controls should be located in the most favorable position for reaching (e.g., lower quarter of panel).

6. Controls used solely for maintenance should be separate from normal operating controls and covered during normal equipment operations.

7. Minimum adjacent edge separation distances for:

a. Rotary selector switch: 1 hand, 1"; 2 hands, 3".

b. Thumbwheel, 0.4".

c. Knob, 1".

d. Pushbuttons, 0.5".

e. Toggleswitch, 0.5".

f. Lever, 1 hand, 2".

g. Pedals, 4".

h. Cranks and handwheels, 3".

8. Minimum and maximum sizes of the following controls should be:

CONTROL	HEIGHT	LENGTH	WIDTH	DIAMETER
a. Rotary selector switch	.625-3"	1-4"	1.0"	
b. Knobs	0.5-1"			1-4"
c. Handwheels				2-4.25" (1 han 7-21" (2 hands
d. Pushbuttons				0.385"-0.75"
e. Toggle switches		0.5-2" (Bare) 1.5-2" (Gloved finger)		

Note. Criteria marked with an asterisk can be evaluated by physically manipulating the controls.

9. Shape coding should be used for rotary controls having dissimilar functions. If size coding is used, use no more than 3 sizes.

10. Controls operated sequentially should be arranged left to right and top to bottom.

*11. Control size and separation distance should permit the use of gloves in the operating environment.

12. "Deadman" controls should be used when the operator's incapacity could produce a critical condition.

13. Extremely critical controls should be protected from inadvertent activation by locating them apart from other controls; by recessing, shielding or covering them; and by using interlocks.

14. Control-display relationships should be apparent through proximity of related controls and displays, through similar groupings, by using a common coding scheme, or by framing or labelling related controls/displays.

*15. Neither the control nor the hand used for operating the control should obscure an associated display.

Note. Criteria marked with an asterisk can be evaluated by physically manipulating the controls.

DISPLAYS

1. Only necessary information should be displayed and only to the required degree of detail.

2. Display face should not be less than 45 degrees to operator's line of sight.

3. Functionally related displays should be located close to each other.

4. Critical and frequently read displays should be located in the center of display panel.

*5. Minimum-maximum viewing distances to display panel: 13-28 inches.

6. Absence of signal should not indicate ON, READY, or OK condition.

7. Displays used while performing maintenance should be covered or non-visible during normal equipment operation.

8. Warning lights should be integrated with or adjacent to remedial controls.

Where possible, the warning light should have a flash rate of 3-5 flashes per second, with approximately equal on-off time.

9. Display colors should read:

a. Steady red for failure or NO GO.

b. Flashing red: emergency.

c. Yellow: marginal.

d. Green: OK.

Avoid blue.

10. Scale markings should be linear and should progress by 1, 2, or 5 units; intermediate marks should not exceed 9.

When positive and negative values are displayed around a zero position, the zero point should be at 12 or 9 o'clock.

11. Pointers should be close to dials to avoid parallax; they should not obscure or exceed width of index marks.

Note. Criteria marked with an asterisk can be evaluated by physically moving the observer. Criterion Number 1 can be evaluated by interviewing operators.

12. Master caution, warning, or summation lights should be set off by themselves.

13. If the face of the scale is color coded, green represents the operating range; yellow, caution; and red, danger.

14. Scales should read clockwise, left to right, or from bottom up; numbers should be oriented upright and placed outside markings.

*15. CRT displays: viewing distance should be approximately 16 inches; they should be hooded or shielded when detection of faint signals is required.

16. Audio displays should be used only when operator vision is overburdened, speech is required, the operator is occupied elsewhere or cannot move his body, or where redundancy or warning signals are required.

17. To check on audio display parameters, see a Human Factors specialist.

18. Audio warning signals: should automatically reset; minimum duration 0.5 second.

19. To check on microphone and headphone parameters, see Human Factors specialist.

20. In high noise environment, use binaural headsets capable of reducing ambient noise to comfortable level.

Note. Criteria marked with an asterisk can be evaluated by physically moving the observer.

LABELS

1. Labels should be placed above controls and displays.

2. Labels should ordinarily be oriented horizontally; vertically only when not critical for personnel safety or performance.

3. Labels should be placed on or near items they identify; ensure that they do not obscure other information and are not obscured by other equipment components.

4. Labels should describe the functions of the units they identify.

5. Labels should be written with standard abbreviations and familiar words only; ensure that trade names are not used.

6. Character height and viewing distance should be in accordance with the following:

VIEWING DISTANCE	HEIGHT
20" or less	0.09"
20-36"	0.17"
36-72"	0.34"
72-144"	0.68"
144-240"	1.13"

7. Conspicuous placards should be placed adjacent to hazardous equipment.

8. Pipe, hose, and tube lines should be clearly labelled as to contents and pressure temperature; electrical receptacles should be marked with voltage, phase, and frequency.

WORKSPACE

1. Displays to be read when standing should be 41-74" above the floor; to be read when seated, between 6-48" above the seat.

2. Controls to be operated when standing should be 34-74" above the floor; to be operated when seated, 8-35" above the seat.

3. Seats should be vertically adjustable from 16-21" in 1" increments. The supporting backrest should be reclinable 103-115 degrees. Arm rests should be provided where feasible.

4. The operator should have at least 4 feet of free space in front of his console.

5. The operator should have the following minimum knee room at his work station: 25" high, 20" wide, 18" deep.

6. Work surface width should be at least 30" wide and 16" deep.

7. Work station table tops should be 29-31" above the floor.

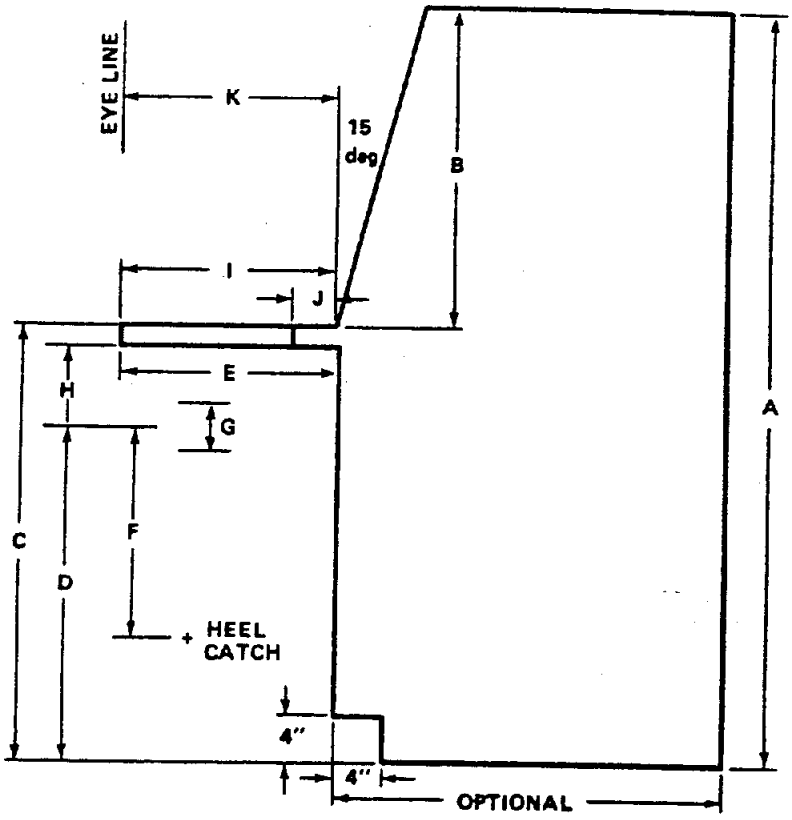
8. Console design should accord with Figure 4-1.

9. Seat design should accord with Figure 4-2.

10. Seats should be large enough to accommodate suitably clothed personnel.

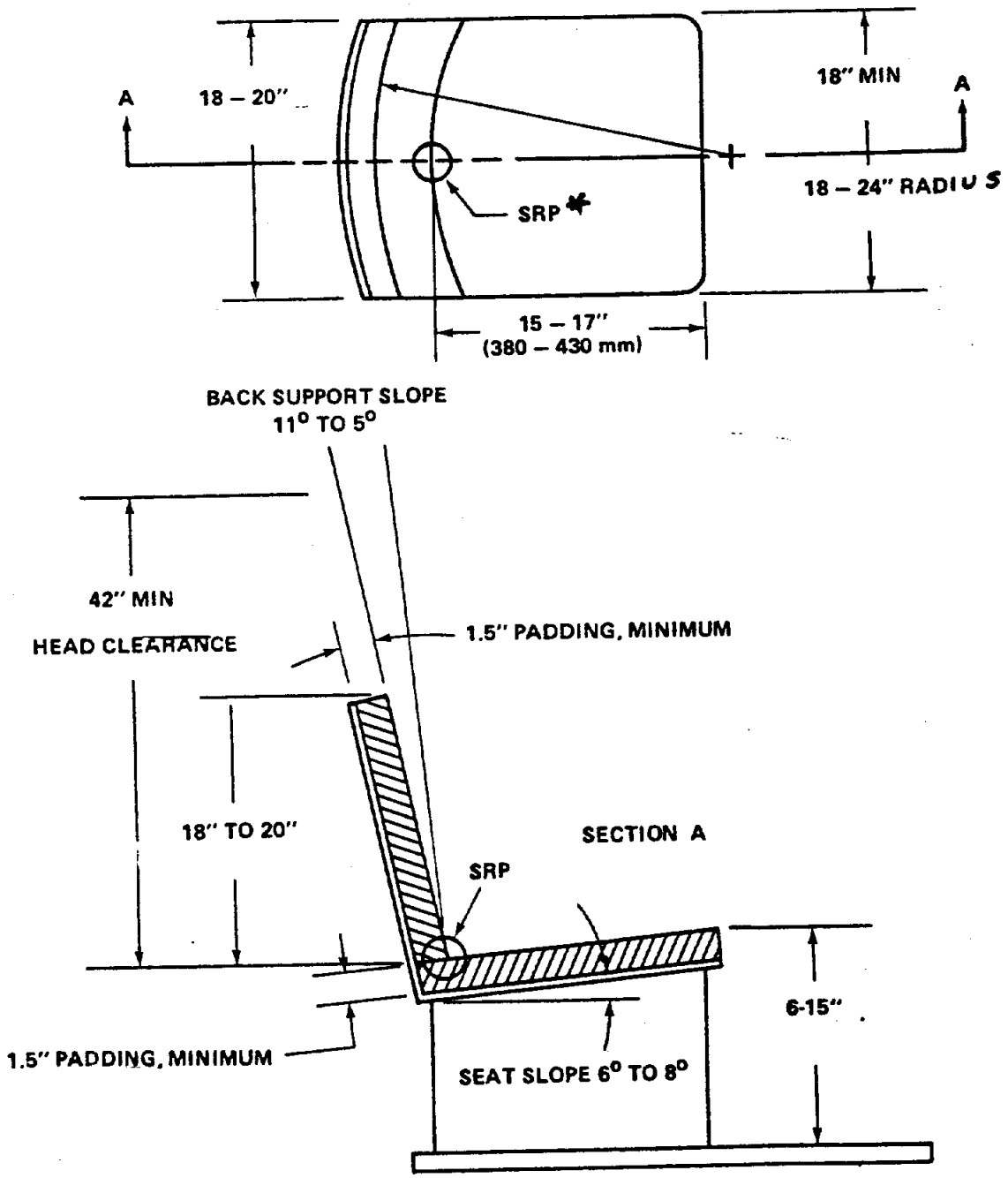
11. The work station should have adequate storage space for paperwork, spare parts, etc.

12. Emergency escape facilities should be provided and marked.



KEY	DIMENSIONS	INCHES
A	MAXIMUM TOTAL CONSOLE HEIGHT FROM STANDING SURFACE	51.5 - 58"
B	SUGGESTED VERTICAL DIMENSION OF PANEL, INCL SILLS	26"
C	WRITING SURFACE: SHELF HEIGHT FROM FLOOR	25.5 - 36"
D	SEAT HEIGHT FROM STANDING SURFACE AT MIDPOINT OF "G"	18 - 28.5"
E	MINIMUM KNEE CLEARANCE	18"
F	FOOT SUPPORT TO SITTING SURFACE	18"
G	SEAT ADJUSTABILITY	5"
H	MINIMUM THIGH CLEARANCE AT MIDPOINT OF "G"	6.5"
I	WRITING SURFACE DEPTH INCLUDING SHELF	16"
J	MINIMUM SHELF DEPTH	4"
K	EYE LINE-TO-CONSOLE FRONT DISTANCE	16"

Figure 4-1. Standard console dimensions key.



* SRP = SEAT REFERENCE POINT

Figure 4-2. Dimensions for vehicle operator's seat.

STAIRS AND LADDERS

1. Fixed ladders should have the following dimensions:

   a. Rung thickness: .75-1.50".

   b. Rung spacing: 9-16".

   c. Ladder width: 12" minimum, 18-21" recommended.

   d. Ladder depth: 30-36".

   e. Space behind ladder: 6" minimum.

   f. Height of ladder above landing: 33" minimun.

2. Toe board or guard screen around platforms should be 3" minimum height.

3. Guard rails should be provided around the opening at the tops of fixed ladders.

4. Adequate footholds should be provided for crew to reach vehicle hatch from ground.

5. Appropriate labels warning of possible hazards should be posted.

6. The exterior ladder tread of personnel platforms should be open grating or nonskid treated.

HANDLES, HANDHOLDS, AND RAILINGS

1. Handles or grasp areas should be located at the equipment's center of gravity and at least 2" from any obstructions.

2. Handles should be recessed where practicable.

*3. Fingers should be able to curve around the handle comfortably.

4. Handles should have minimum 2" clearance during use.

5. Stairs, ladders, platforms, and ramps should be equipped with a handrail on each side.

*6. Handholds should be furnished where needed and within easy reach; they should accommodate gloved hands.

7. The handrail diameter should be 1.25-3".

8. Guardrails for personnel platforms should have a top rail height of not less than 42". The distance between the platform edge and the centerline of the railing should not exceed 2.5".

9. Handrail to handrail width on stair ladders should be 21-24"; on one-way stairs, 30" minimum; on two-way stairs, 48" minimum.

10. Handrail clearance from the wall should be a minimum of 1.75" for stairs; for stair ladders, 2".

11. Handrails should have nonslip surfaces.

Note. Criteria marked with an asterisk can be evaluated by physically performing required actions.

DOORS, HATCHES, AND ENTRYWAYS

1. Wall hatches should be flush with the floor where structurally possible.

2. Round hatches should have a minimum 30" diameter.

3. Rectangular hatches to be used by personnel with light clothing should be (for top and bottom access) 23" wide, 13" high; for side access, 30" wide, 26" high. If used by personnel with bulky clothing, they should be (for top and bottom access) 27" wide, 16" high; for side access they should be 34" wide, 29" high.

4. Floor escape hatches should have a minimum 22" diameter.

5. Latch handles should release down or forward; catch and lock, up or aft. They should be operable with gloved hands.

*6. Emergency doors should be quick opening.

*7. Escape openings should be free of obstructions and permit passage of personnel with survival gear.

Note. Criteria marked with an asterisk can be evaluated by physically performing required actions.

1. Temperature in a moving vehicle should be 50 degrees (F) minimum, 85 degrees (F) maximum. Above the maximum temperature, air conditioning should be provided.

2. Noise level should permit necessary person-to-person and telephone communication without shouting. If specific measures of noise level are required, see paragraph 5.8.3 and following of MIL-STD 1472B and refer to a Human Engineering specialist.

3. If specific measures of humidity are required, see paragraph 5.8.1.4 of MIL-STD 1472B and refer to Human Engineering specialist.

4. If specific measures of illumination are required, see Table XIX, pp. 141-144 of MIL-STD 1472B, and refer to Human Engineering specialist.

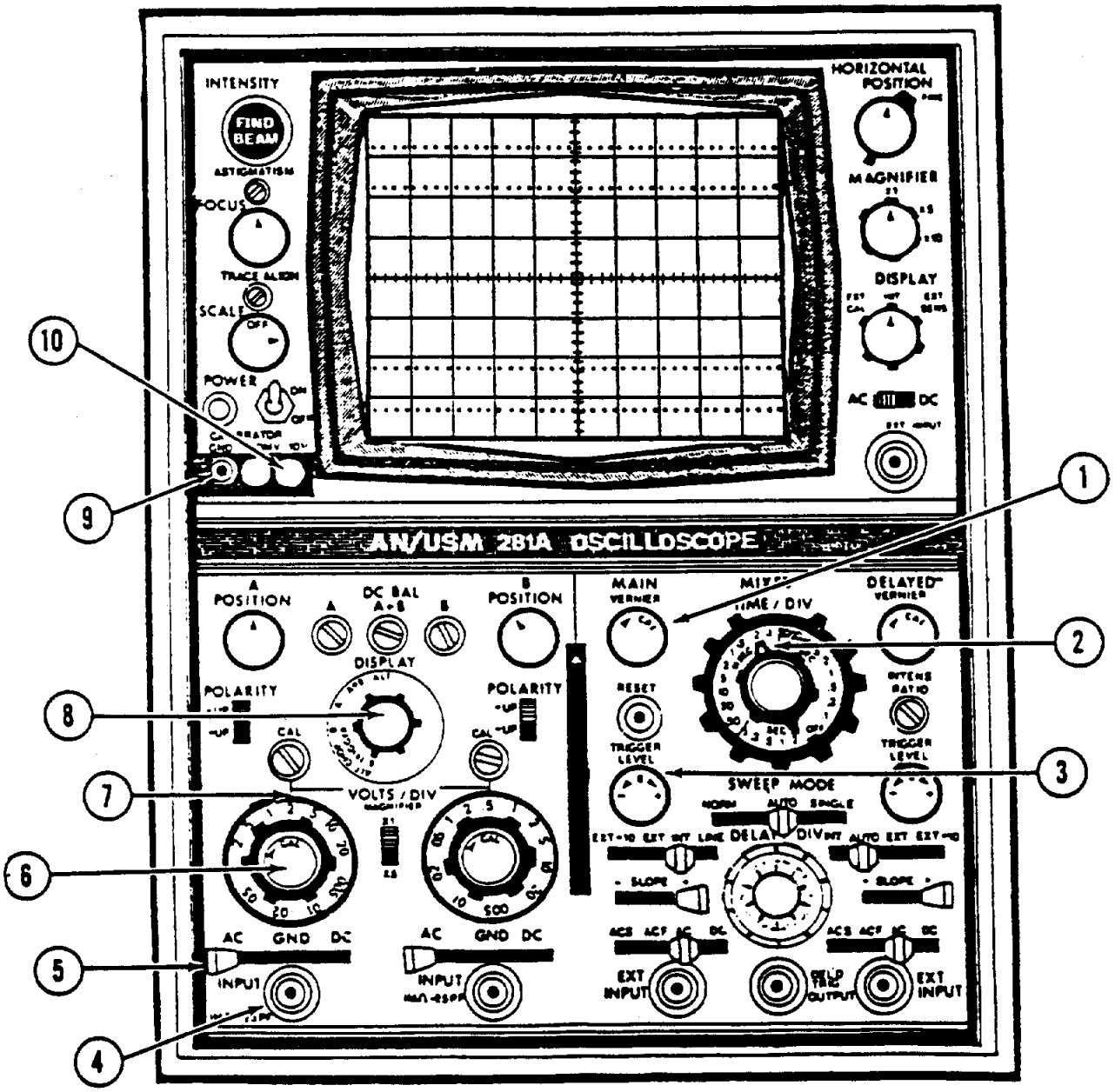
5. If specific measures of vibration are required, see Figure 36, pp. 149 of MIL-STD 1472B, and refer to Human Engineering specialist.

6. The lateral field of view in a moving vehicle should be 180 degrees minimum. The vehicle operator should be able to view the ground at all distances beyond 10 feet in front of the vehicle. Upward visibility should be 15 degrees above the horizontal.

Note. The adequacy of the work environment can also be ascertained by interviewing personnel and/or by asking them to rate the adequacy of the environment. For rating scale development, see Section Three of this report.

Example of the Simulated Observation Form

Adjunct to Structured Observation Form



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

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

"What make or brand is the newest one?"

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

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

Mayer and Piper Format After Modification

"What make or brand is the newest one?"

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

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

Comparisons of Questionnaire Layout

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

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

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

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

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

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

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

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

Conclusions Regarding Questionnaire Layout

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

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

Description of Branching

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

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

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

Examples of Branching

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

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

ASK EVERYONE:

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

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

Comparisons of Branching

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

Conclusions Regarding Branching

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

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

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

CHAPTER VIII

FUTURE RESEARCH

Introduction

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

Background Issues

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

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

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

Emerging Relevant Technologies

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

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

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

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

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

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

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

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

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

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

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

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

Primary Research Recommendations

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

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

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

Selection Rationale for Research Recommendations

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

- Scale Development Procedures and Analyses

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

- Procedural Guides to Item Wording

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

- Subjective Workload Assessment Methods

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

- Automated Portable Test Systems

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

- Cognitive Complexity

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

- Behaviorally Anchored Rating Scales (BARS)

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

- Item Nonresponse, Branching, and Demographic Characteristics

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

- Pictorial Anchors

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

BIBLIOGRAPHY

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Bernardin, H. J., & Boetcher, R. (1978, August). The effects of rater training and cognitive complexity on psychometric error in ratings. Paper presented at the meeting of the American Psychological Association, Toronto.
Cognitive Complexity

Bernardin, H. J., Cardy, R. L., & Carlyle, J. J. (1982). Cognitive complexity and appraisal effectiveness: Back to the drawing board? Journal of Applied Psychology, 67(2), 151-160.
Cognitive Complexity

Bernardin, H. J., & Kane, J. S. (1980). A second look at behavioral observation scales. Personnel Psychology, 33, 809-814.
Behavioral Observation Scales

Bernardin, H. J., La Shells, M. B., Smith, P. C., & Alvares, K. M. (1976, February). Behavioral expectation scales: Effects of developmental procedures and formats. Journal of Applied Psychology, 61(1), 75-79.
Behavioral Expectation Scales
Behaviorally Anchored Rating Scales
Questionnaire Layout

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65, 60-66.

Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored rating scales. Journal of Applied Psychology, 66(4), 458-463.
Behaviorally Anchored Rating Scales

Bernardin, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 62(1), 64-69.
Behavioral Expectation Scales
Behaviorally Anchored Rating Scales

Bickley, W. R. (1980, September). Training device effectiveness: Formulation and evaluation of methodology (Report No. RR 1291). U.S. Army Research Institute for the Behavioral and Social Sciences.
Multiple-Choice Scales (AD A122 777)

Bieri, J. (1966). Cognitive complexity and personality development. In O. J. Harvey (Ed.), Experience, structure, and adaptability. New York: Springer.

Biggs, J. B. (1970). Personality correlates of some dimensions of study behavior. Australian Journal of Psychology, 22, 287-297.
"Don't Know" Category

Birnbaum, M. H. (1981, March). Reason to avoid triangular designs in nonmetric scaling. Perception and Psychophysics, 29(3), 291-293.

Bittner, A. C., Carter, R. C., & Krause, M. (1981, November). Performance tests for repeated measures: Moran and computer batteries (NBDL 81R012). New Orleans, LA: U.S. Naval Biodynamics Lab.

Blackburn, R. S. (1982, Spring). Multidimensional scaling and the organizational sciences. Journal of Management, 8(1), 95-103.

Blanz, F., & Ghselli, E. E. (1972). This mixed-standard scale: A new rating system. Personnel Psychology, 25, 185-199.
Cognitive Complexity
Mixed Standard Scales

Block, A. S., & Jouett, M. L. (1978, June). Development of a performance evaluation tool for on-the-job competency appraisal for respiratory therapists -- Subproject III (HRP-0901397). Dallas, TX: American Association for Respiratory Therapy.
Multiple-Choice Scales

Blower, D. J. (1980, July). The bias in the presentation of stimuli when the up and down method is used with forced choice responding (NAMRL-1269). Pensacola, FL: Naval Aerospace Medical Research Laboratory. (DTIC No. AD A089790)

Blower, D. J. (1981, August). Determining visual acuity thresholds: A simulation study of stimulus presentation strategies (NAMRL-1269). Pensacola, FL: Naval Aerospace Medical Research Laboratory, Naval Air Station. (DTIC No. AD A111821)
Multiple-Choice Scales

Boote, A. S. (1981). Reliability testing of psychographic scales. Journal of Advertising Research, 21(5), 53-60.
Number of Scale Points
Response Alternatives

Bordeleau, Y., & Turgeon, B. (1977). Comparison of 3 psychometric methods used in attitude questionnaires. Canadian Journal of Behavioural Science, 9(1), 26-36.

Borman, W. C. (1975). Effects of instruction to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 60, 556-560.
Behaviorally Anchored Rating Scales

Borman, W. C. (1977, December). Consistency of rating accuracy and rating errors in the judgment of human performance. Organizational Behavior and Human Performance, 20(2), 238-252.

Borman, W. C. (1979). Format and training effects on rater accuracy and rater errors. Journal of Applied Psychology, 64, 410-421.
Behaviorally Anchored Rating Scales
Cognitive Complexity
Questionnaire Layout

Borman, W. C. (1981, June). Performance ratings: Comments on the state of the art. In Cecil J. Mullins, AFHRL Conference on Human Appraisal Proceedings. Reviewed and submitted for publication by Lonnie D. Valentin, Jr., Chief, Force Acquisition Branch, Manpower and Personnel Division, Air Force Human Resources Laboratory, AFHRL Technical Paper, 81-20.

Borman, W. C., & Dunnette, M. (1975). Behavior based versus task-oriented performance ratings: An empirical study. Journal of Applied Psychology, 60, 561-565.
Behavioral Expectation Scales
Response Alternatives

Borman, W. C., & Rosse, R. L. (1980, September). Peer ratings: Scoring strategy development and reliability demonstration on Air Force basic trainees (AFHRL-TR-80-28). Minneapolis, MN: Personnel Decisions Research Institute. (DTIC No. AD A090325)

Bradburn, N. M., & Sudman, S. (1979). Improving interview method and questionnaire design (3rd printing). San Francisco, CA: Jossey-Bass Publishers.
Age
Branching
Ethnic Background
Interviewing
Length of Items and Number of Items
Open-Ended Items and Closed-End Items

Bradburn, N. M., Sudman, S., Blair, E., & Stocking, C. (1978, Summer). Question threat and response bias. Public Opinion Quarterly, 42(2), 221-234.

Bradley, R. A. (1982, May). Paired comparisons (FSU Statistics Report No. M615; ONR Technical Report No. 157). Tallahassee, FL: The Florida State University, Department of Statistics. (DTIC No. AD A123877)
Paired-Comparison Items

Brannon, R. (1981). Current methodological issues in paper-and-pencil measuring instruments. Psychology of Women Quarterly, 5(4), 618-627.
Gender
Multiple-Choice Scales

Brown, Curtis A. (1982, May). The effect of factor range on weight and scale values in a linear averaging model. Dissertation Abstracts International, 42(11-B).

Bruvold, W. H. (1977). Reconciliation of apparent nonequivalence among alternative rating methods. Journal of Applied Psychology, 62(1), 111-115.

Burns, A. C., & Harrison, C. (1979). A test of the reliability of psychographics. Journal of Marketing Research, 16, 32-38.

Butler, M. C., & Jones, A. P. (1979). The health opinion survey reconsidered: Dimensionality, reliability, and validity. Journal of Clinical Psychology, 35(3), 554-559.

Carter, R. C., Kennedy, R. S., Bittner, A. C., & Krause, M. (1981, July). Item recognition as a performance evaluation test for environmental research. New Orleans, LA: U.S. Naval Biodynamics Laboratory.

Carter, R. C., & Sbisa, H. E. (1982, January). Human performance tests for repeated measurements: Alternate forms of eight tests by computer (NBDL-82R003). New Orleans, LA: U.S. Naval Biodynamics Lab.

Carter, R. C., Stone, D. A., & Bittner, A. C. (1982). Repeated measurements of manual dexterity applications and support of the two-process theory. Ergonomics, 25(9), 829-838.

Cascio, W. F., & Valenzi, E. R. (1977). Behaviorally anchored rating scores: Effects of education and job experience of raters and ratees. Journal of Applied Psychology, 62, 278-282.
Education

Chapman, R. G., & Staelin, R. (1982, August). Exploiting rank ordered choice set data within the stochastic utility model. Journal of Marketing Research, 19, 288-301.

Checklist: How effective is your management of personnel (1978, September). Focus on Employee Relations, by the Bank Personnel Division staff of the American Bankers Associations.

Christian, J. K., & Bringmann, W. G. (1982). Comparison of computerized versus standardized feedback and accurate versus inaccurate feedback. Psychological Reports, 50, 1067-1070.

Chun, K., Fields, V., & Friedman, S. (1975, August). Military attitudinal surveys: An overview. In H. W. Sinaiko, & L. A. Broedling (Eds.), Perspectives on attitude assessment: Surveys and their alternatives. Manpower Research and Advisory Services, Smithsonian Institution, prepared under the Navy Manpower R&D Program of the Office of Naval Research, N00014.67-A-0399.0006.
Future Research
Interviewing

Church, F. (1983, June). Questionnaire construction manual for operational tests and evaluation. Prepared for the Deputy Commander of Tactics and Test, 57th Fighter Weapons Wing/DT, Tactical Fighter Weapons Center (TFWC), Nellis AFB, NV.
Interviewing
Number of Scale Points
Rank Order Scales
Semantic Differential Scales

Cicchinelli, L. F., Harmon, K. R., & Keller, R. A. (1982, December). Relative cost and training effectiveness of the 6883 F-111 converter/flight control system simulators as compared to actual equipment (AFHRL-TR-82-30). Lowry AFB, CO: Logistics and Technical Training Division.
Multiple-Choice Scales
Open-Ended Items and Closed-End Items

Cocanougher, A. B., & Ivancevich, J. M. (1978, July). "BARS" performance rating for sales force personnel. Journal of Marketing, 87-95.
Behaviorally Anchored Rating Scales

Cole, N. (1973). On measuring the vocational interests of women. Journal of Counseling Psychology, 20, 105-112.
Continuous and Circular Scales

Comrey, A. L., & Montag, I. (1982, Summer). Comparison of factor analytic results with two-choice and seven-choice personality item formats. Applied Psychological Measurement, 6(3), 285-289.
Number of Scale Points

Cooper, W. H. (1981, September). Ubiquitous halo. Psychological Bulletin, 90(2), 218-244.

Cottle, C. E., & McKeown, B. (1980, January). The forced-free distinction in Q technique: A note on unused categories in the Q sort continuum. Operant Subjectivity, 3(2), 58-63.

Couch, A., & Keniston, K. (1960). Yea sayers and nay sayers: Agreeing response set as a personality variable. Journal of Abnormal and Social Psychology, 60, 151-174.
Balanced Items

Craig, C. S., & McCann, J. M. (1978). Item nonresponse in mail surveys: Extent and correlation. Journal of Marketing Research, 15, 285-289.
Age

Cronbach, L. J. (1946). Response sets and test validity. Educational and Psychological Measurement, 6, 475-494.
Balanced Items

Cronbach, L. J. (1950). Further evidence on response sets and test design. Educational and Psychological Measurement, 10, 3-31.
Balanced Items

Crooks, L. A. (Ed.). (1972). An investigation of sources of bias in the prediction of job performance: A six year study. Princeton, NJ: Educational Testing Service.
Ethnic Background

Cudeck, R. (1980, Summer). A comparative study of indices for internal consistency. Journal of Educational Measurement, 17(2), 117-130.

Dambrot, F. (1980, April). Test item order and academic ability, or should you shuffle the test item deck? Teaching of Psychology, 7(2), 94-96.
Order of Items

Daniel, F. R., Jr., & Wagner, E. E. (1982). Differences among Holland types as measured by the hand test: An attempt at construct validation. Educational and Psychological Measurement, 42, 1295-1301.

Daniel, W. W., Schott, B., Atkins, F. C., & Davis, A. (1982, Spring). An adjustment for nonresponse in sample surveys. Educational and Psychological Measurement, 42(1), 57-67.

Deaton, W. L., Glasnapp, D. R., & Poggio, J. P. (1980). Effects of item characteristics on psychometric properties of forced choice scales. Educational and Psychological Measurement, 40, 599-610.
Balanced Items
Multiple-Choice Scales
Wording of Items and Tone of Wording

DeCotiis, T. A. (1977). An analysis of the external validity and applied relevance of three rating formats. Organizational Behavior and Human Performance, 19, 247-266.

DeCotiis, T. A. (1978). A critique and suggested revision of behaviorally anchored rating scales developmental procedures. Educational and Psychological Measurement, 38, 681-690.

De Jung, J. E., & Kaplan, H. (1962). Some differential effects of race of rater and combat attitude. Journal of Applied Psychology, 46, 370-374.
Ethnic Background

Devine, P. J. (1980, November). An investigation of the degree of correspondence among four methods of item bias analysis. Dissertation Abstracts International, 41(5-B).

Dickinson, T. L., & Zellinger, P. M. (1980). A comparison of behaviorally anchored rating and mixed standard scale formats. Journal of Applied Psychology, 65(2), 147-154.
Mixed Standard Scales

Dickson, J., & Albaum, G. (1977). A method for developing tailormade semantic differentials for specific marketing content areas. Journal of Marketing Research, 14, 87-91.
Bipolar Scales
Multiple-Choice Scales
Semantic Differential Scales

Divgi, D. R. (1981). Model-free evaluation of equating and scaling. Applied Psychological Measurement, 5(2), 203-208.

Dixon, D. J., Copeland, M. G., & Halcomb, C. G. (1978, May). Psychomotor battery approaches to performance prediction and evaluation in hyperbaric, thermal and vibratory environments: Annotated bibliographies and integrative review (NBDL-M002). Lubbock, TX: Texas Tech University. (DTIC No. AD A099981)

Dolch, N. A. (1980). Attitude measurement by semantic differential on a bipolar scale. The Journal of Psychology, 105, 151-154.
Bipolar Scales
Response Alternatives
Semantic Differential Scales

Dollard, J. A., Dixon, M. L., & McCann, P. H. (1980, September). Shipboard instruction and training management with computer technology: A pilot application (NPRDC-TR-80-34). San Diego, CA: Navy Personnel Research and Development Center.
Multiple-Choice Scales
Middle Scale Point Position

Downs, P. E. (1978). Miscellany: Testing the upgraded semantic differential. Journal of the Market Research Society, 20(2), 99-103.
Semantic Differential Scales

Downs, S., Farr, R. M., & Colbeck, L. (1978). Self-appraisal: A convergence of selection and guidance. Journal of Occupational Psychology, 51, 271-278.

Drasgow, F. (1982). Biased test items and differential validity. Psychological Bulletin, 92(2), 526-531.

Drasgow, F. (1982, Summer). Choice of test model for appropriateness measurement. Applied Psychological Measurement, 6(3), 297-308.

Drasgow, F., & Miller, H. E. (1982). Psychometric and substantive issues in scale construction and validation. Journal of Applied Psychology, 67(3), 268-279.

Dziuban, C. D., & Shirkey, E. C. (1980). Sampling adequacy and the semantic differential. Psychological Reports, 47, 351-357.
Semantic Differential Scales

Eckblad, G. (1980). The curvex: Simple order structure revealed in ratings of complexity, interestingness, and pleasantness. Scandinavian Journal of Psychology, 21(1), 1-16.

Edvardsson, B. (1980). Effect of reversal of response scales in a questionnaire. Perceptual and Motor Skills, 50, 1125-1126.

Edwards, M. R. (1981, August). Improving performance appraisal by using multiple appraisers. Industrial Management and Data Systems, 13-16.
Paired-Comparison Items

Edwards, R. H. (1981). Coefficients of effective length. Educational and Psychological Measurement, 41, 283-285.

Eggemeier, F. T., Crabtree, M. S., & La Point, P. A. (1983, October). The effect of delayed report on subjective ratings of mental workload. Proceedings of the Human Factors Society 27th Annual Meeting, 139-143.
Future Research

Eggemeier, F. T., Crabtree, M. S., Zingg, J. J., Reid, G. B., & Shingledecker, C. A. (1982). Subjective workload assessment in a memory update task. Proceedings of the Human Factors Society 26th Annual Meeting, 643-647.
Future Research

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

Eiser, J. R., & Osmon, B. E. (1978). Judgmental perspective and value connotations of response scale labels. Journal of Personality and Social Psychology, 36(5), 491-497.
Balanced Items
Bipolar Scales

Eiser, J. R., & Stroebe, W. (1972). Categorization and social judgment. London: Academic Press.

Eisler, H. (1982). On the nature of subjective scales. Scandinavian Journal of Psychology, 23(3), 161-171.

Elithorn, A., Mornington, S., & Stavron, A. (1982). Automated psychological testing: Some principles and practice. International Journal of Man-Machine Studies, 17(3), 247-263.

Evaluating performance: A self-instruction unit (HRP-0902012), (1979). Princeton, NJ: Educational Testing Service.

Fay, C. H., & Latham, G. P. (1982). Effects of training and rating scales on rating errors. Personnel Psychology, 35, 105-116.
Behavioral Expectation Scales
Behavioral Observation Scales

Ferber, R. (1966). Item nonresponse in a consumer survey. Public Opinion Quarterly, 30, 399-415.
Age

Finley, D. M. (1976, May). The effects of scale continuity and behavioral anchor specificity upon the psychometric properties of performance rating scales. Dissertation Abstracts International, 36(11-B).

Finley, D. M., Osborn, H. G., Dubin, J. A., & Jeanneret, P. R. (1977). Behaviorally based rating scales: Effects of specific anchors and disguised scale continua. Personnel Psychology, 30, 659-669.
Mixed Standard Scales

Finn, R. H. (1976). Concerning questionnaire quality and operational utility. Catalog of Selected Documents in Psychology, 6, 32.

Fischer, G. H., & Formann, A. K. (1982, Fall). Some applications of logistic latent trait models with linear constraints on the parameters. Applied Psychological Measurement, 6(4), 397-416.

Fivars, G. (1975). The critical incident technique: A bibliography. JSAS Catalog of Selected Documents in Psychology, 5, 210.
Behavioral Observation Scales

Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-1358.
Behavioral Observation Scales

Fleishmann, U. (1981). Psychometric techniques and questionnaires for use in gerontopsychological investigations. International Journal of Rehabilitation Research, 4(1), 96-97.

Ford, D. L. (1976, February). Predicting group decision strategies: The effect of rating-scale use bias on accuracy of prediction. Catalog of Selected Documents in Psychology, 6(3).

Fowler, F. J. (1984). Survey research methods. Beverly Hills, CA: Sage Publications.

Fowler, F. J., & Mangione, T. W. (1983). The role of interviewing training and supervision in reducing interviewer effects on survey data. Proceedings of the American Statistical Association Meeting, Survey Research Methods Section, 124-128.

Fralicx, R. D., & Raju, N. S. (1982). A comparison of five methods for combining multiple criteria into a single composite. Educational and Psychological Measurement, 42, 823-827.

Frankel, L. R. (1975, August). Restrictions to survey sampling legal, practical, and ethical. In H. W. Sinaiko, & L. A. Broedling (Eds.), Perspectives on attitude assessment: Surveys and their alternatives. Manpower Research and Advisory Services, Smithsonian Institution, prepared under the Navy Manpower R&D Program of the Office of Naval Research (N00014.67-A-0399.0006).
Future Research

Frederiksen, N., Jensen, O., & Beaton, A. E. (1972). Prediction of organizational behavior. New York: Pergamon Press, Inc.
Semantic Differential Scales

Friedman, B. A., & Cornelius, E. T. (1976). Effect of rater participation in scale construction on the psychometric characteristics of two rating scale formats. Journal of Applied Psychology, 61(2), 210-216.
Behaviorally Anchored Rating Scales

Fullerton, J. T., & Holley, M. (1982). A new look at standard scale transformation. Psychological Reports, 50, 1148-1150.

Furnham, A., & Henderson, M. (1982). The good, the bad and the mad: Response bias in self-report measures. Personality and Individual Differences, 3(3), 311-320.

Garner, W. R. (1960). Rating scales, discriminability, and information transmission. Psychological Review, 67, 343-352.
Number of Scale Points

Gay, L. R. (1980). The comparative effects of multiple-choice versus short-answer tests on retention. Journal of Educational Measurement, 17(1), 45-50.

Gerow, J. R. (1980, April). Performance on achievement tests as a function of the order of item difficulty. Teaching of Psychology, 7(2), 93-94.
Order of Items

Gibbons, J. D., Olkin, I., & Sobel, M. (1979, September). A subset selection technique of scoring items on a multiple choice test. Psychometrika, 44(3), 259-270.

Gividen, G. M. (1973, February). Order of merit: Descriptive phrases for questionnaires. Unpublished report, available from the ARI Field Unit at Fort Hood, TX.
Middle Scale Point Position

Glaser, M. A., & Collen, M. F. (1972). Toward automated medical decisions. Computers and Biomedical Research, 5, 180-189.
Future Research

Goodstadt, M. S., & Magid, S. (1977). When Thurstone and Likert agree -- A confounding of methodologies. Educational and Psychological Measurements, 37, 811-818.

Graef, J., & Spence, I. (1979, January). Using distance information in the design of large multidimensional scaling experiments. Psychological Bulletin, 86(1), 60-66.

Green, B. F. (1981). A primer of testing. American Psychologist, 36(10), 1001-1011.
Multiple-Choice Scales

Groves, R. M. (1979). Actors and questions in telephone and personal interview surveys. Public Opinion Quarterly, 43, 190-205.
Interviewing
Response Alternatives

Guilford, J. P. (1954). Psychometric methods. New York: McGraw-Hill.
Number of Scale Points

Guion, R. M. (1979, April). Principles of work sample testing. II. A non-empirical taxonomy of test uses (TR-79-A8). Bowling Green, OH: Bowling Green State University. (DTIC No. AD A072446)
Number of Scale Points

Guion, R. M., & Ironson, G. H. (1983, February). Latent trait theory for organizational research. Organizational Behavior and Human Performance, 31(1), 54-87.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stauffer, et al. (Eds.), Measurement and prediction. Princeton, NJ: Princeton University Press.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. Lazarfeld (Ed.), Mathematical thinking in the social sciences. Glencoe, IL: Free Press.
Continuous and Circular Scales

Guttman, L. (1957). Empirical verification of the radex structure of mental abilities and personality traits. Educational and Psychological Measurement, 17, 391-407.
Continuous and Circular Scales

Hahn, J. E. (1981, March). A Monte Carlo study of incomplete data designs and configurations in non-metric multidimensional scaling. Dissertation Abstracts International, 41(9-B).

Hambleton, R. K., Mills, C. N., & Simon, R. (1983). Determining the lengths for criterion referenced tests. Journal of Educational Measurement, 20(1), 27-38.

Hambleton, R. K., & van der Linden, W. J. (1982, Fall). Advances in item response theory and applications: An introduction. Applied Psychological Measurement, 6(4), 373-378.

Hamel, C. J., Braby, R., Terrell, W. R., & Thomas, G. (1983, January). Effectiveness of job training materials based on three format models: A field evaluation (Technical Report 138). Orlando, FL: Training Analysis and Evaluation Group, Department of the Navy.
Multiple-Choice Scales

Hamelink, J., & Hamelink, J. (1980). A numeric plan for performance appraisal. Training and Development Journal, 34(10), 88-89.

Hammer, W. C., Kim, J. S., Baird, L., & Bigoness, N. J. (1974). Race and sex as determinants of ratings by potential employers in a simulated work sampling task. Journal of Applied Psychology, 59, 705-711.
Ethnic Background

Hardt, R. H., Eyde, L. D., Primoff, E. S., & Tordy, G. R. (1978). The New York State Trooper job element examination. Albany, NY: New York Police.
Ethnic Background

Hartke, A. R. (1979, Fall). The development of conceptually independent sub-scales in the measurement of attitudes. Educational and Psychological Measurement, 39(3), 585-592.

Harvey, R. J. (1982, April). The future of partial correlation as a means to reduce halo in performance ratings. Journal of Applied Psychology, 67(2), 171-176.

Hatchett, S., & Schuman, H. (1975). White respondents and race-of-interviewer effects. Public Opinion Quarterly, 39, 523-528.
Interviewing

Heavlin, W. D., Lee-Merrow, S. W., & Lewis, V. M. (1982, Fall). The psychometric foundations of goal attainment scaling. Community Mental Health Journal, 18(3), 230-241.

Helm, W. R., & Donnell, M. L. (1979). Mission operability assessment technique: A system evaluation methodology (TP 79-31). Point Mugu, CA: Pacific Missile Test Center. (NTIS No. ADB042746)
Multiple-Choice Scales

Hendel, D. D. (1979). Paired comparisons intransitivity: Is it relatively stable over time? Educational and Psychological Measurement, 39, 779-784.

Hirschman, E. C., & Wallendorf, M. R. (1982). Free-response and card-sort techniques for assessing cognitive content: Two studies concerning their stability, validity, and utility. Perceptual and Motor Skills, 54, 1095-1110.

Holbrook, M. B. (1977). Comparing multiattribute attitude models by optimal scaling. Journal of Consumer Research, 4(3), 165-171.

Holland, P. W. (1981, March). When are item response models consistent with observed data? Psychometrika, 46(1), 79-92.

Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. Journal of Applied Psychology, 63(5), 579-588.

Hom, P. W., De Nisi, A. S., Kinicki, A. J., & Bannister, B. D. (1982). Effectiveness of performance feedback from behaviorally anchored rating scales. Journal of Applied Psychology, 67(5), 568-576.
Behaviorally Anchored Rating Scales

Horayangkura, V. (1978, December). Semantic dimensional structures: A methodological approach. Environment and Behavior, 10(4), 555-584.

Hsu, L. M. (1979). A comparison of three methods of scoring true-false tests. Educational and Psychological Measurement, 39, 785-790.

Hughes, G. D. (1975). Upgrading the semantic differential. Journal of the Marketing Research Society, 17(1), 41-44.
Semantic Differential Scales

Hulin, C. L. (1982, April). Some reflections on general performance dimensions and halo rating error. Journal of Applied Psychology, 67(2), 165-170.

Hulin, C. L., Drasgow, F., & Komocar, J. (1980, July). Applications of item response theory to analysis of attitude scale translations (80-5). Champaign, IL: Department of Psychology, University of Illinois. (DTIC No. AD A087834)

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. Journal of Applied Psychology, 67(6), 818-825.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982, Summer). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. Applied Psychological Measurement, 6(3), 249-260.

Humphreys, M. A. (1982, Winter). Data collection effects on nonmetric multidimensional scaling solutions. Educational and Psychological Measurement, 42(4), 1005-1022.

Hyman, H. H., Cobb, W. J., Feldman, J. J., Hart, C. W., & Stember, C. H. (1954). Interviewing in social research. Chicago: University of Chicago Press.
Ethnic Background
Interviewing

Imada, A. S., & London, M. (1979). Relationship between subjects, scales, and stimuli in research on social perceptions. Perceptual and Motor Skills, 48, 691-697.
Ethnic Background

Innes, J. M. (1977). Extremity and "don't know" sets in questionnaire response. British Journal of Social Clinical Psychology, 16, 9-12.
"Don't Know" Category

Ironson, G. H., & Smith, P. C. (1981, Summer). Anchors away -- the stability of meaning of anchors when their location is changed. Personnel Psychology, 34(2), 249-262.

Ivancevich, J. M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. Journal of Applied Psychology, 64(5), 502-508.
Behavioral Expectation Scales

Ivancevich, J. M. (1980). Behavioral expectation scales versus nonanchored and trait rating systems: A sales personnel application. Applied Psychological Measurement, 4(1), 131-133.
Behavioral Expectation Scales
Response Alternatives

Jacobs, R., Kafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. Personnel Psychology, 33, 595-640.
Behaviorally Anchored Rating Scales
Cognitive Complexity

Jensen, H. E., Massey, I. H., & Valentine, L. D. (1977). Armed Services Vocational Aptitude Battery Development (ASVAB Forms 5, 6 and 7) (Technical Research Note 77-33). Fort Sheridan, IL: U.S. Army Enlisted Processing Command.

Jesteadt, W. (1980). An adaptive procedure for subjective judgments. Perception and Psychophysics, 28, 85-88.
Multiple-Choice Scales

Johnson, J. D. (1981). Effects of the order of presentation of evaluative dimensions for bipolar scales in four societies. The Journal of Social Psychology, 113, 21-27.
Balanced Items
Bipolar Scales

Jones, A. P., Main, D. S., Butler, M. C., & Johnson, L. A. (1982). Narrative job descriptions as potential sources of job analysis ratings. Personnel Psychology, 35, 813-828.

Kafry, D., Zedeck, S., & Jacobs, R. (1976). Short notes: The scalability of behavioral expectation scales as a function of developmental criteria. Journal of Applied Psychology, 61(4), 519-522.
Behavioral Expectation Scales

Kalton, G. (1983). Introduction to survey sampling. Beverly Hills, CA: Sage Publications.

Kane, J. S., & Bernardin, H. J. (1982). Behavioral observation scales and the evaluation of performance appraisal effectiveness. Personnel Psychology, 35, 635-641.
Behavioral Observation Scales

Kane, J. S., & Lawler, E. E. (1979). Performance appraisal effectiveness: Its assessment and determinants. In B. Staw (Ed.), Research in organizational behavior (Vol. 1) (pp. 425-478). Greenwich, CT: JAI Press.

Katcher, B. L., & Bartlett, C. J. (1979, April). Rating errors of inconsistency as a function of dimensionality of behavioral anchors (Research Report No. 84). College Park, MD: University of Maryland, Department of Psychology. (DTIC No. AD A068922)
Mixed Standard Scales

Kearney, W. J. (1976, June). The value of behaviorally based performance appraisals. Business Horizons, 75-83.

Kearney, W. J. (1979, January). Behaviorally anchored rating scales -- MBO's missing ingredient. Personnel Journal, 58(1), 20-25.
Behaviorally Anchored Rating Scales
Multiple-Choice Scales

Keaveny, T. J., & McCann, A. F. (1975). A comparison of behavioral expectation scales and graphic rating scales. Journal of Applied Psychology, 60, 695-703.

Kelly, G. A. (1955). The psychology of personal constructs (Vol. 1). New York: Norton Press.
Cognitive Complexity

Kennedy, R., Bittner, A., Jr., Carter, R., Krause, M., Harbeson, M., McCafferty, D., Pepper, R., & Wiker, S. (1981, July). Performance evaluation tests for environmental research (PETER): Collected papers (NBDL-80R008). New Orleans, LA: Naval Biodynamics Laboratory. (DTIC No. AD A111296)

Kennedy, R. S., Bittner, A. C., Jr., Harbeson, M. M., & Jones, M. B. (1981, November). Perspectives in performance evaluation tests for environmental research (PETER) (NBDL-803004). New Orleans, LA: Naval Biodynamics Laboratory. (DTIC No. AD A111180)

Kennedy, R. S., Jones, M. B., & Harbeson, M. M. (1981, November). Assessing productivity and well-being in Navy workplaces. New Orleans, LA: U.S. Naval Biodynamics Laboratory.

Kesselman, G., Lopez, F. M., & Lopez, F. E. (1982). The development and validation of self-report scored in-basket test in an assessment center setting. Public Personnel Management, 11(3), 228-238.

Kiesler, C. A., Collins, B. E., & Miller, N. (1969). Attitude change: A critical analysis of theoretical approaches. New York: John Wiley & Sons, Inc.
Semantic Differential Scales

Kingstrom, P. O. (1979). The effects of rater-ratee interactions and the format of appraisal interviews on rating characteristics and feedback. Dissertation Abstracts, 39(10-B), 5114.
Interviewing

Kingstrom, P. O., & Bass, A. R. (1981). A critical analysis of studies comparing behaviorally anchored rating scales (BARS) and other rating formats. Personnel Psychology, 34, 263-289.
Behaviorally Anchored Rating Scales

Klimoski, R. J., & London, M. (1974). Role of rater in performance appraisal. Journal of Applied Psychology, 59(4), 445-451.

Klockars, A. J. (1979). Evaluative confounding in the choice of bipolar scales. Psychological Report, 45, 771-775.
Balanced Items
Bipolar Scales
Semantic Differential Scales

Klockars, A. J., King, D. W., & King, L. A. (1981). The dimensionality of bipolar scales in self-description. Applied Psychological Measurement, 5(2), 219-227.
Bipolar Scales
Semantic Differential Scales

Klores, M. S. (1966). Rater bias in forced-distribution ratings. Personnel Psychology, 19, 411-421.
Age

Koch, W. R., & Reckase, M. D. (1978, June). A live tailored testing comparison study of the one- and three-parameter logistic models (Research Report 78-1). Columbia, MO: Department of Educational Psychology, University of Missouri. (DTIC No. AD A058528)

Koenig, R. (1983, April 18). Interest rises in testing by computer. The Wall Street Journal.
Future Research

Korschot, B. C. (1978, July-August). Quantitative evaluation of investment research analysts. Financial Analysts Journal, 41-46.

Kowalski, R. (1984). AI and software engineering. Datamation, 30(18), 92-102.

Krohn, G. S. (1984). LAVM/RV OT II human factors assessment materials. Fort Hood, TX: U.S. Army Research Institute for the Behavioral and Social Sciences (ARI), HQ, TCATA, PERI-SH.
Interviewing
Number of Scale Points

Krus, D. J., & Krus, P. H. (1977, Spring). Normal scaling of the unidimensional dominance matrices: The domain referenced model. Educational and Psychological Measurement, 37(1), 189-193.

Labaw, P. (1980). Advanced questionnaire design (2nd ed.). Cambridge, MA: Abt Books.
Branching
Interviewing
Order of Items
Questionnaire Layout
Wording of Items and Tone of Wording

Lahey, M. A., & Saal, F. E. (1981). Evidence incompatible with a cognitive compatibility theory of rating behavior. Journal of Applied Psychology, 66(6), 706-715.
Cognitive Complexity

Lamont, L. M., & Lundstrom, W. J. (1977, November). Identifying successful industrial salesmen by personality and personal characteristics. Journal of Marketing Research, 14, 517-529.

Lampert, S. I. (1979). The Attitude Pollimeter: A new attitude scaling device. Journal of Marketing Research, 16, 578-582.
Age
Continuous and Circular Scales
Education

Landy, F., & Farr, J. (1980). Performance rating. Psychological Bulletin, 87, 72-107.
Age
Behaviorally Anchored Rating Scales
Education
Ethnic Background
Gender
Response Alternatives

Landy, F. J., & Barnes, J. L. (1979). Scaling behavioral anchors. Applied Psychological Measurement, 3(2), 193-200.
Behaviorally Anchored Rating Scales
Paired-Comparison Scales

Landy, F. J., Barnes-Farrell, J. L., Vance, R. J., & Steele, J. W. (1980). Statistical control of halo error in performance ratings. Journal of Applied Psychology, 65(5), 501-506.

Landy, F. J., Farr, J. L., Saal, F. E., & Freytag, W. R. (1976). Behaviorally anchored scales for rating the performance of police officers. Journal of Applied Psychology, 61(6), 750-758.
Behaviorally Anchored Rating Scales

Landy, F. J., Vance, R. J., & Barnes-Farrell, J. L. (1982, April). Statistical control of halo: A response. Journal of Applied Psychology, 67(2), 177-180.

La Rocco, J. M., & Butler, M. C. (1977, December). Survey questionnaires: More than meets the eye (77-57). San Diego, CA: Naval Health Research Center. (DTIC No. AD A100262)

Latham, G. P., Fay, C. H., & Saari, L. M. (1979). The development of behavioral observation scales for appraising the performance of foremen. Personnel Psychology, 32, 299-311.
Behavioral Expectation Scales
Behavioral Observation Scales

Latham, G. P., Saari, L. M., & Fay, C. (1980). BOS, BES, and baloney: Raising kane with Bernardin. Personnel Psychology, 33, 815-821.

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. Personnel Psychology, 30, 255-268.
Behavioral Observation Scales

Latham, G. P., & Wexley, K. N. (1981). Increasing productivity through performance appraisal. Reading, MA: Addison-Wesley.

Layne, B. H., & Thompson, D. N. (1981). Questionnaire page length and return rate. The Journal of Social Psychology, 113, 291-292.
Length of Items and Number of Items
Questionnaire Layout

Lee, R., Malone, M., & Greco, S. (1981). Multitrait-multimethod-multirater analysis of performance ratings for law enforcement personnel. Journal of Applied Psychology, 66(5), 625-632.

Lee, R., Miller, K. J., & Graham, W. K. (1982). Corrections for restriction of range and attenuation in criterion-related validation studies. Journal of Applied Psychology, 67(5), 637-639.

Levine, E. L., Flory, A., & Ash, R. A. (1977). Self-assignment in personnel selection. Journal of Applied Psychology, 62(4), 428-435.
Ethnic Background

Lienert, G. A., & Raatz, U. (1981, Spring). Item homogeneity defined by multivariate symmetry. Applied Psychological Measurement, 5(2), 263-269.

Link, S. W. (1982, September). Correcting response measures for guessing and partial information. Psychological Bulletin, 92(2), 469-486.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. Journal of Applied Psychology, 60, 10-13.
Number of Scale Points

Locander, W. B., & Staples, W. A. (1978). Evaluating and motivating salesmen with the BARS method. Industrial Marketing Management, 7, 43-48.
Behaviorally Anchored Rating Scales

Lodge, M. (1981). Magnitude scaling: Quantitative measurement of opinions. Sage University Paper Series: Quantitative Applications in the Social Sciences, 25.

London, M., & Poplowski, J. R. (1976). Effects of information on stereotypic development in performance appraisal and interview contexts. Journal of Applied Psychology, 61(2), 199-205.
Gender

Lord, F. M. (1982, Fall). Standard error of an equating by item response theory. Applied Psychological Measurement, 6(4), 463-472.

Lund, T. (1975, April). An alternative content method for multidimensional scaling. Multivariate Behavioral Research, 10(2), 181-191.

Mabe, P. A., III, & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. Journal of Applied Psychology, 67(3), 280-296.

Malhotra, N. A. (1981). A scale to measure self-concepts, person concepts, and product concepts. Journal of Marketing Research, 15, 456-464.
Future Research
Semantic Differential Scales

Mann, I. T., Phillips, J. L., & Thompson, E. G. (1979). An examination of methodological issues relevant to the use and interpretation of the semantic differential. Applied Psychological Measurement, 3, 213-229.
Bipolar Scales
Semantic Differential Scales

Marascuilo, L. A., & Slaughter, R. E. (1981, Winter). Statistical procedures for identifying possible sources of item bias based on -sub (x) -sup-2 statistics. Journal of Educational Measurement, 18(4), 229-248.

Marshall, S. P. (1981). Sequential item selection: Optimal and heuristic policies. Journal of Mathematical Psychology, 23, 134-152.

Martin, W. S. (1978, May). Effects of scaling on the correlation coefficient: Additional considerations. Journal of Marketing Research, 15, 304-308.

Martins, G. R. (1984). The overselling of expert systems. Datamation, 30(18), 76-80.
Future Research

Massey, R. H., Mullings, C. J., & Earles, J. A. (1978, December). Performance appraisal ratings: The content issue (AFHRL TR-78-69). Brooks Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory. (DTIC No. AD A064690)

Matarazzo, J. D. (1983). Computerized psychological testing. Science, 221, (4608).
Future Research

Mathews, J. L., Wright, C. E., Yudowitch, K. L., Geddie, J. C., & Palmer, R. L. (1978, August). The perceived favorableness of selected scale anchors and response alternatives (Technical Paper 319). Palo Alto, CA: Operations Research Associates, and Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (DTIC No. AD A061755)
Middle Scale Point Position
Response Alternatives

Maul, T. L., & Pargman, D. (1975). The semantic differential as a psychometric instrument for behavioral research in sport. International Journal of Sport Psychology, 9(1), 7-15.
Semantic Differential Scales

Maurelli, V. A., & Weiss, D. J. (1981, November). Factors influencing the psychometric charracteristics of an adaptive testing strategy for test batteries (Research Report 81-4). Minneapolis, MN: Department of Psychology, University of Minnesota. (DTIC No. AD A109666)

Mayer, C. S., & Piper, C. (1982). A note on the importance of layout in self-administered questionnaires. Journal of Marketing Research, 19(3), 390-391.
Questionnaire Layout

Mayerberg, C. K., & Bean, A. G. (1978, Fall). Two types of factors in the analysis of semantic differential attitude data. Applied Psychological Measurement, 2(4), 469-480.

McBride, J. R., Sympson, J. B., Vale, C. D., Pine, S. M., & Bejar, I. I. (1977, March). Applications of computerized adaptive testing (Research Report 77-1). Minneapolis, MN: Department of Psychology, University of Minnesota. (DTIC No. AD A038114)

McCormick, C. C., & Kavanagh, J. A. (1981). Scaling interpersonal checklist items to a circular model. Applied Psychological Measurement, 5(4), 421-447.
Continuous and Circular Scales

McDonald, R. P. (1981). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 34(1), 100-117.

McDonald, R. P. (1982, Fall). Linear versus nonlinear models in item response theory. Applied Psychological Measurement, 6(4), 379-396.

McFarland, S. G. (1981). Effects of question order on survey responses. Public Opinion Quarterly, 45, 208-215.
Gender
Order of Items

McIver, J. P., & Carmines, E. G. (1981). Unidimensional scaling. Sage University Paper series on quantitative applications in the social sciences, 07-024. Beverly Hills and London: Sage Publishers.
Behavioral Expectation Scales
Behavioral Observation Scales
Mixed Standard Scales
Multiple-Choice Scales
Paired-Comparison Items
Rank Order Scales

McKelvie, S. J. (1978). Graphic rating scales -- how many categories? British Journal of Psychology, 69, 185-202.
Continuous and Circular Scales
Number of Scale Points
Semantic Differential Scales

Meisels, M., & Ford, L. H. (1969). Social desirability response set and semantic differential judgments. Journal of Social Psychology, 78, 45-54.
Gender

Mellenbergh, G. J. (1982, Summer). Contingency table models for assessing item bias. Journal of Educational Statistics, 7(2), 105-118.

Melzer, C. W., Koeslag, J. H., & Schach, S. S. (1981). Correction of item-test correlations and attempts at improving reproducibility in item-analysis: An experimental approach. Educational and Psychological Measurement, 41(4), 979-990.

Menezes, D., & Elbert, N. F. (1979). Alternative semantic scaling formats for measuring store image: An evaluation. Journal of Marketing Research, 16(1), 80-87.
Questionnaire Layout
Response Alternatives

Meriwether, T. N. (1979, May). Developing a job analysis based performance appraisal system at the United States Military Academy: A new approach to forced-choice evaluation. Dissertation Abstracts International, 39(11-B).

Messick, D. M., & Van de Geer, J. P. (1981, November). A reversal paradox. Psychological Bulletin, 90(3), 582-593.

Messmer, D. J., & Seymour, D. T. (1982, Summer). The effects of branching in item response. Public Opinion Quarterly, 46(2), 270-277.
Age
Branching
Education

Meyer, H. H. (1980). Self-appraisal of job performance. Personnel Psychology, 33, 291-295.

Miller, P. M. (1974). A note in sex differences on the sementic differentials. British Journal of Social Clinical Psychology, 13, 33-36.
Gender

Mokken, R. J., & Lewis, C. (1982, Fall). A nonparametric approach to the analysis of dichotomous item responses. Applied Psychological Measurement, 6(4), 417-430.

Mooney, G. R. (1981, September). An analysis of rater error in relation to ratee job characteristics. Dissertation Abstracts International, 42(3-B).

Moroney, W. F. (1984). The use of checklists and questionnaires during system and equipment test and evaluation. Shrivenham, England: NATO Defense Research Group Panel VIII Workshop, Applications of Systems Ergonomics to Weapon System Development, Royal Military College of Science, Vol 1, C-59-C-68.
Future Research
Multiple-Choice Scales

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

Mueller, J. E. (1973). War, presidents and public opinion. New York: John Wiley & Sons.
Wording of Items and Tone of Wording

Mullins, C. J. (1981, June). AFHRL conference on human appraisal: Proceedings (AFHRL-TP-81-20). Brooks Air Force Base, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. (DTIC No. AD A102755)

Mullins, C. J., Earles, J. A., & Wilbourn, J. M. (1979, May). Personnel rating effectiveness as a function of number of rating statements (AFHRL-TR-79-11). Brooks Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory.
Length of Items and Number of Items

Munson, J. M., & McIntyre, S. H. (1979). Developing practical procedures for the measurement of personal values in cross-cultural marketing. Journal of Marketing Research, 16, 48-52.
Paired-Comparison Scales

Murphy, J. W. (1980). Use of behaviorally anchored rating scales (BARS) to complement the management by objectives (MBO) and fitness report components of the Marine Corps performance evaluation system. Master of Military Art and Sciences (MMAS) thesis prepared at U.S. Army Command and General Staff College, Fort Leavenworth, KS. (DTIC No. AD A097694)
Behaviorally Anchored Rating Scales

Murphy, K., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. (1982). Relationship between observational accuracy and accuracy in evaluating performance. Journal of Applied Psychology, 67, 320-325.

Murphy, K. R. (1982, April). Difficulties in the statistical control of halo. Journal of Applied Psychology, 67(2), 161-164.

Murphy, K. R., Martin, C., & Garcia, M. (1982). Do behavioral observation scales measure observation? Journal of Applied Psychology, 67(5), 562-567.
Behavioral Observation Scales

Myczyk, J. P., & Gable, M. (1981). Unidimensional (global) vs. multidimensional composite performance appraisals of store managers. Journal of Academic Marketing Science, 9(3), 191-205.

Myers, E. (1984). Business takes the fifth. Datamation, 30(195), 53-57.
Future Research

Neill, J. A., & Jackson, D. N. (1976, Spring). Minimum redundancy item analysis. Educational and Psychological Measurement, 36(1), 123-134.

Nemeroff, W. F., & Wexley, K. N. (1979). An exploration of the relationships between performance feedback interview characteristics and interview outcomes as perceived by managers and subordinates. Journal of Occupational Psychology, 52. 25-34.
Interviewing

Nevo, B. (1980, Summer). Item analysis with small samples. Applied Psychological Measurement, 4(3), 323-329.

Ng, S. H. (1982). Choosing between the ranking and rating procedures for the comparison of values across cultures. European Journal of Social Psychology, 12(2), 169-172.

Norton, S. D., Gustafson, D. P., & Foster, C. E. (1977). Assessment for management potential: Scale design and development, training effects and rater/ratee sex effects. Academy of Management Journal, 20, 117-131.
Gender

Nugent, W. A., Laabs, G. J., & Panell, R. C. (1982, February). Performance test objectivity: A comparison of rater accuracy and reliability using three observation forms (NPRDC TR 82-30). San Diego, CA: Navy Personnel Research and Development Center. (DTIC No. AD A111077)
Questionnaire Layout

Null, C. H. (1980, April). Design considerations for multidimensional scaling. Behavior Research Methods and Instrumentation, 12(2), 274-280.

Nunnally, J. C. (1967). Psychometric theory. New York: McGraw-Hill.
Number of Scale Points

Nunnally, J. C., Leonard, L. C., & Wilson, W. H. (1977). Studies of voluntary visual attention -- theory, methods, and psychometric issues. Applied Psychological Measurement, 1, 203-218.

Nygren, T. E. (1982, April). Conjoint measurement and conjoint scaling: A users guide (AFAMRL-TR-82-22). Wright-Patterson Air Force Base, OH: Air Force Aerospace Medical Research Laboratory, Aerospace Medical Division.

Oaster, T. R. F. (1982). Index for differentiation of state and trait scales with new stressor scales. Psychological Reports, 51, 272.

Oborne, D. J. (1976). Examples of the use of rating scales in ergonomics research. Applied Ergonomics, 7(4), 201-204.
Continuous and Circular Scales

O'Neil, M. J. (1979). Estimating the non-response bias due to refusals in telephone surveys. Public Opinion Quarterly, 43, 218-232.
Age
Education
Ethnic Background

Orlich, D. C. (1978). Designing sensible surveys. Pleasantville, NY: Redgrave Publishing Company.
Behavioral Observation Scales
Interviewing
Multiple-Choice Scales
Open-Ended Items and Closed-End Items
Rank Order Scales
Wording of Items and Tone of Wording

Orpen, C. (1981). The effect of examiner ethnicity on the job satisfaction responses of Blacks in community surveys: A South African study. Journal of Community Psychology, 9(1), 81-85.

Ory, J. C. (1982). Item placement and wording effects on overall ratings. Educational and Psychological Measurement, 42, 767-775.
Balanced Items
Wording of Items and Tone of Wording

Osgood, C. E., Suci, G., & Tannenbaum, P. (1957). The measurement of meaning. Urbana, IL: University of Illinois Press.
Bipolar Scales
Gender
Semantic Differential Scales

Palachek, A. D., & Kerin, R. A. (1980, August). Alternative approaches to the two-group concordance problem in brand preference rankings. Journal of Marketing Research, 386-389.

Paradise, L. V., & Kottler, J. (1979, August). Use of Q-factor analysis for initial instrument validation. Psychological Reports, 45(1), 139-143.

Parkinson, R. (1977, November). Recipe for a realistic appraisal system. Personnel Management, 37-40.

Parsons, C. K., & Hulin, C. L. (1982). An empirical comparison of item response theory and hierarchical factor analysis in applications to the measurement of job satisfaction. Journal of Applied Psychology, 67(6), 826-834.

Peabody, D. (1967). Trait inferences: Evaluative and descriptive aspects. Journal of Personality and Social Psychology Monograph, 7, pp. 644.

Peterson, J. M. (1977, August). The influence of favorable context on questionnaire response. Dissertation Abstracts International, 381(2-A), 747.

Petz, B., & Mayer, D. (1977). Comparison between Thurstone's Law of Comparative Judgments scale and the Sum of Totals scale. Acta Instituti Psychologici, No. 79-86, 55-64.

Presser, S., & Schuman, H. (1980). The measurement of a middle position in attitude surveys. Public Opinion Quarterly, 44(1), 70-85.
Education
Middle Scale Point Position

Primoff, E. S. (1978, August). The use of self-assessments in examining. A paper presented at the 86th Annual Convention of the American Psychological Association, Toronto, Ontario, Canada. (NTIS No. PB-298 606)

Pulakos, E. D. (1984, May). The development of training programs to increase accuracy with different rating formats (84-2). East Lansing, MI: Department of Psychology, Michigan State University.

Pursell, E. D., Campion, M. A., & Gaylord, S. R. (1980, November). Structured interviewing: Avoiding selection problems. Personnel Journal, 907-912.

Pursell, E. D., Dossett, D. L., & Latham, G. P. (1980). Obtaining valid predictors by minimizing rating errors in the criterion. Personnel Psychology, 33, 91-97.

Range, L. M., Anderson, H. N., & Wesley, A. L. (1982, October). Personality correlates of multiple choice answer-changing patterns. Psychological Reports, 51(2), 523-527.

Ray, J. J. (1980). The comparative validity of Likert, projective, and forced-choice indices of achievement motivation. The Journal of Social Psychology, 111, 63-72.
Multiple-Choice Scales

Ray, J. J. (1982). The construct validity of balanced Likert scales. The Journal of Social Psychology, 118, 141-142.
Balanced Items

Ray, J. J., & Bozek, R. S. (1979, June). NSCALE II: A program to analyze and score multiscale surveys and test batteries. Behavior Research Methods and Instrumentation, 11(3), 402.

Reid, G. B., Eggemeier, F. T., & Nygren, T. E. (1982). An individual differences approach to SWAT scale development. Proceedings of the Human Factors Society 26th Annual Meeting, 639-642.
Future Research

Reid, G. B., Shingledecker, C. A., & Eggemeier, F. T. (1981). Application of conjoint measurement to workload scale development. Proceedings of the Human Factors Society 25th Annual Meeting, 522-526.
Future Research

Reid, G. B., Shingledecker, C. A., Nygren, T. E., & Eggemeier, F. T. (1981, October). Development of multidimensional subjective measures of workload. Proceedings of the International Conference on Cybernetics and Society, sponsored by IEEE Systems, Man and Cybernetics Society, Atlanta, GA, 403-406.
Future Research

Resnick, S. M., & Mohrman, A. M. (1982, October). The design of performance appraisal systems: Some implication from research findings (G 8205-924). Los Angeles, CA: University of Southern California, Graduate School of Business Administration. (DTIC No. AD A120333)

Reynolds, T. J. (1976, June). The analysis of dominance matrices: Extraction of unidimensional orders within a multidimensional context (Technical Report No. 3). Los Angeles, CA: Department of Psychology, University of Southern California. (DTIC No. AD A029450)

Reynolds, T. J. (1981, Fall). ERGO: A new approach to multidimensional item analysis. Educational and Psychological Measurement, 41(3), 643-659.

Reynolds, T. J., & Jolly, J. P. (1980). Measuring personal values: An evaluation of alternative methods. Journal of Marketing Research, 17, 531-536.
Future Research
Paired-Comparison Items
Rank Order Scales
Response Alternatives

Ridgway, J., MacCullough, M. J., & Mills, H. E. (1982). Some experiences in administering a psychometric test with a light pen and microcomputer. International Journal of Man-Machine Studies, 17(3), 265-278.

Rigney, J. W., Towne, D. M., Moran, P. J., & Mishler, R. A. (1980, July). Field evaluation of the generalized maintenance trainer-simulator, II. AN/SPA-66 Radar Repeater (NPRDC TR-8-30-2). San Diego, CA: Navy Personnel Research and Development Center.
Rank Order Scales

Rizzo, W. A., & Frank, F. D. (1977, Fall). Influence of irrelevant cues and alternate forms of graphic rating scales on the halo effect. Personnel Psychology, 30(3), 405-417.

Robertson, I. T., & Kandola, R. S. (1982). Work sample tests: Validity, adverse impact and applicant reaction. Journal of Occupational Psychology, 55, 171-183.

Rokeach, M. (1973). The nature of values. New York: Free Press.
Paired-Comparison Items

Roscoe, J. T. (1975). Fundamental research statistics for the behavioral sciences. New York: Holt, Rinehart, and Winston, Inc.
Multiple-Choice Scales

Rose, G. L. (1978). Sex effects on effort attributions in managerial performance evaluation. Organizational Behavior and Human Performance, 21, 367-378.
Gender

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

Rounds, J. B., Jr., Miller, T. W., & Dawis, R. V. (1978, Summer). Comparability of multiple rank order and paired comparison methods. Applied Psychological Measurement, 2(3), 415-422.

Roy, J. J. (1982). Machieavellianism, forced-choice formats and the validity of the F scale: A rejoinder to Bloom. Journal of Clinical Psychology, 38(4), 779-782.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980, Spring). A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 17(1), 1-10.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980, Fall). Biased item detection techniques. Journal of Educational Statistics, 5(2), 213-233.

Ryans, A. B., & Srinivasan, V. (1979, November). Improved method for comparing rank-order preferences of two groups of consumers. Journal of Marketing Research, 16, 583-587.

Saal, F. E. (1979). Mixed standard rating scale: A consistent system for numerically coding inconsistent response combinations. Journal of Applied Psychology, 64(4), 422-428.
Cognitive Complexity
Mixed Standard Scales

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88(2), 413-428.

Saito, T. (1977, July). Multidimensional Thurstonian scaling with an application to color metrics. Japanese Psychological Research, 19(2), 78-89.

Samejima, F. (1979, February). Constant information model: A new, promising item characteristic function (Research Report 79-1). Knoxville, TN: Department of Psychology, University of Tennessee. (DTIC No. AD A070090)

Samejima, F. (1979, December). A new family of models for the multiple-choice item (Research Report 79-4). Knoxville, TN: Department of Psychology, University of Tennessee. (DTIC No. AD A080350)

Sauser, W. I. (1979). A comparative evaluation of the effects of rater participation and rater training on characteristics of employee performance appraisal ratings and related mediating variables. Dissertation Abstracts International, 39 (10-B), 5116.

Sauser, W. I., & Pond, S. B. (1981). Effects of rater training and participation of cognitive complexity: An exploration of Schneier's cognitive reinterpretation. Personnel Psychology, 34, 563-577.
Cognitive Complexity

Schaefer, B. A., Bavelas, J., & Bavelas, A. (1980). Using echo technique to construct student-generated faculty evaluation questionnaires. Teaching of Psychology, 7(2), 83-86.
Wording of Items and Tone of Wording

Schaefer, E. S. (1961). Converging conceptual models for maternal behavior and for child behavior. In J. Glidewell (Ed.), Parental attitudes and child behavior. Springfield, IL: Thomas.
Continuous and Circular Scales

Schaeffer, N. (1980). Evaluating race-of-interviewer effects in a national survey. Sociological Methods and Research, 8, 400-419.
Ethnic Background
Interviewing

Schein, V. E. (1973). The relationship between sex role stereotypes and requisite management characteristics. Journal of Applied Psychology, 57, 95-100.
Gender

Schertzer, C. B. (1982). Semantic properties of commonly used scaling adjectives. Dissertation Abstracts International, 42(8-A), 3733.

Schmidt, F. L., & Hunter, J. E. (1980, September). Differential and simple-group validity of employment tests by race: A critical analysis of three recent studies. Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development Center, Research Branch.
Ethnic Background

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1982). Progress in validity generalization: Comments on Callender and Osburn and further developments. Journal of Applied Psychology, 67(6), 835-845.

Schmidt, F. L., & Johnson, R. H. (1973). Effect of race on peer ratings in an industrial setting. Journal of Applied Psychology, 57, 237-241.
Ethnic Background

Schmitt, J. C., & Scheirer, C. J. (1977). The effect of item order on objective tests. Teaching of Psychology, 4, 144-145.
Order of Items

Schneier, C. E. (1977a). Operational utility and psychometric characteristics of behavioral expectation scales: A cognitive reinterpretation. Journal of Applied Psychology, 62(5), 541-548.
Behavioral Expectation Scales
Behaviorally Anchored Rating Scales
Cognitive Complexity
Mixed Standard Scales

Schneier, C. E. (1977b). Multiple rater groups and performance appraisal. Public Personnel Management, 6(1), 13-20.

Schneier, C. E., & Beatty, R. W. (1979a). Integrating behaviorally-based and effectiveness-based methods. The Personnel Administrator, 65-76.
Behaviorally Anchored Rating Scales

Schneier, C. E., & Beatty, R. W. (1979b). Developing behaviorally anchored rating scales. The Personnel Administrator, 59-68.
Behaviorally Anchored Rating Scales

Schneier, C. E., & Beatty, R. W. (1979c). Combining BARS and MBO: Using an appraisal system to diagnose performance problems. The Personnel Administrator, 51-60.
Behaviorally Anchored Rating Scales

Schonemann, P. H. (1982, June). A metric for bounded response scales. Bulletin of the Psychonomic Society, 19(6), 317-319.

Schriesheim, C. A. (1981). The effect of grouping or randomized items on leniency response bias. Educational and Psychological Measurement, 41, 401-411.

Schriesheim, C. A. (1981). Leniency effects on convergent and discriminant validity for grouped questionnaire items: A further investigation. Educational and Psychological Measurement, 41, 1093-1099.

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. Educational and Psychological Measurement, 41, 1101-1114.

Schroder, H. M. (1971). Conceptual complexity and personality organization. In H. M. Schroder & P. Suedfeld (Eds.), Personality theory and information processing. New York: Ronald.

Schuman, H., & Converse, J. (1971). The effects of black and white interviewers in black responses in 1968. Public Opinion Quarterly, 35, 44-68.
Ethnic Background
Interviewing

Schuman, H., & Presser, S. (1981). Questions and answers in attitude surveys: Experiments on question form, wording, and context. New York: Academic Press, Inc.
Age
Balanced Items
"Don't Know" Category
Education
Ethnic Background
Gender
Middle Scale Point Position
Multiple-Choice Scales
Open-Ended Items and Closed-End Items
Order of Items
Wording of Items and Tone of Wording

Schutz, H. G., & Rucker, M. H. (1975). A comparison of variable configurations across scale lengths: An empirical study. Organizational and Psychological Measurement, 35, 319-324.
Number of Scale Points

Schwab, D. P., Heneman, H. H., & De Cotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 28, 549-562.

Scott, W. A. (1968). Attitude measurement. In G. Lindzey, & E. Aronson (Eds.), The handbook of social psychology. Reading, MA: Addison-Wesley.

Sedlacek, W. E. (1977). Test bias and the elimination of racism. Journal of College Student Personnel, 18(1), 16-20.
Ethnic Background

Segal, D. R., & Savell, J. M. (1975, August). Research on race relations in the U.S. Army: The multi-method matrix. In H. W. Sinciko, & L. A. Broedling (Eds.), Perspectives on attitude assessment: Surveys and their alternatives. Manpower Research and Advisory Services, Smithsonian Institution.
Ethnic Background

Shanker, P. (1977). Scaling trait-words by cross modality watching. Indian Psychological Review, 15(1), 1-6.

Shannon, R. H. (1981a). The validity of task analytic information to human performance research in unusual environments. In R. H. Shannon, & R. C. Carter (Eds.), Task analysis and the ability requirements of tasks: Collected Papers (NBDL-81R009). New Orleans, LA: Naval Biodynamics Laboratory.
Multiple-Choice Scales

Shannon, R. H. (1981b). Performance evaluation tests for environmental research (PETER) using task analysis. In R. H. Shannon, & R. C. Carter (Eds.), Task analysis and the ability requirements of tasks: Collected papers (NBDL-81R0009). New Orleans, LA: Naval Biodynamics Laboratory.
Rank Order Scales

Shannon, R. H. (1981c). The utility of task analytic techniques to research in unusual environments. In R. H. Shannon, & R. C. Carter (Eds.), Task analysis and the ability requirements of tasks: Collected papers (NBDL-81R009). New Orleans, LA: Naval Biodynamics Laboratory.
Rank Order Scales

Shannon, R. H. (1981d). A factor analytic approach to determining stability of human performance (NBDL-81R010). New Orleans, LA: U.S. Naval Biodynamics Laboratory.

Shannon, R. H. (1981e). Task analytic approach to human performance battery development. New Orleans, LA: U.S. Naval Biodynamics Laboratory.

Shannon, R. H., & Carter, R. C. (1981, September). Task analysis and the ability requirements of tasks: Collected papers (NBDL-81R009). New Orleans, LA: Naval Biodynamics Laboratory. (DTIC No. AD A111181)
Rank Order Scales

Sharon, A. T. (1980, May). An investigation of reference ratings of applicants for administrative law judge (PRR-80-6). Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development Center. (NTIS No. PB80-203573)

Sheehy, G. (1981). Pathfinders. New York: William Morrow and Company, Inc.
Length of Items and Number of Items

Shingledecker, C. A. (1983, June). Behavioral and subjective workload metrics for operational environments. AGARD-AMP Symposium: Sustained Intensive Air Operations: Physiological and Performance Aspects, Paris.
Future Research
Multiple-Choice Scales

Shosteck, H., & Fairweather, W. R. (1979). Physician response rates to mail and personal interviews surveys. Public Opinion Quarterly, 43, 206-217.
Interviewing

Silverstein, A. B. (1980, Summer). Item intercorrelations, item-test correlations, and test reliability. Educational and Psychological Measurement, 40(2), 353-355.

Sinaiko, H. W., & Broedling, L. A. (1975, August). Perspectives on attitude assessment: Surveys and their alternatives - Proceedings of a conference (TR-2). Washington, DC: Smithsonian Institution.

Singh, A. C., & Bilsbury, C. D. (1982). Scaling subjective variables by SPC (sequential pair comparisons). Behavioural Psychotherapy, 10(2), 128-145.

Sitton, L. R., Adams, I. G., & Anderson, H. N. (1980). Personality correlates of students' patterns of changing answers on multiple-choice tests. Psychological Reports, 47, 655-660.

Slater, P. E. (1962). Parent behavior and the personality of the child. Journal of Genetic Psychology, 101, 53-68.
Continuous and Circular Scales

Smith, K. F., & Baldauf, R. B., Jr. (1982). The concurrent validity of self-rating with interviewer rating in the Australian second language proficiency ratings scale. Educational and Psychological Measurement, 42, 1117-1124.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.
Behavioral Expectation Scales
Behaviorally Anchored Rating Scales

Smith, T. E. (1976, December). Scalable choice models. Journal of Mathematical Psychology, 14(3), 239-243.

Smith, T. W. (1981). Qualifications to generalize absolutes: "Approval of hitting" questions on the GSS. Public Opinion Quarterly, 45, 224-230.
Education
Gender
Wording of Items and Tone of Wording

Soeken, K. L., & Macready, G. B. (1982, September). Respondents' perceived protection when using randomized response. Psychological Bulletin, 92(2), 487-489.

Space, L. G. (1981). The computer as a psychometrician. Behavior Research Methods & Instrumentation, 13(4), 595-606.
Future Research

Spector, P. E. (1976). Choosing response categories for summated rating scales. Journal of Applied Psychology, 61(3), 374-375.

Spiers, P. A, & Pihl, R. O. (1976). The effect of study habits, personality, and order of presentation on success in an open-book objective examination. Teaching of Psychology, 3, 33-34.
Order of Items

Spies-Wood, E. (1980). Learned helplessness and item difficulty ordering. Psychologia Africana, 29-40.
Order of Items

Stager, P., & Paine, T. G. (1980). Separation discrimination in a simulated air traffic control display. Human Factors, 22(5), 631-636.

Staples, W. B., & Locander, W. B. (1975). Behaviorally anchored scales: A new tool for retail management evaluation and control. Journal of Retailing, 52(4), 39-95.
Behaviorally Anchored Rating Scales

Steinheiser, F. H., Jr., Epstein, K. I., Mirabella, A., & Macready, G. B. (1978, August). Criterion-referenced testing: A critical analysis of selected models (Technical Paper 306). College Park, MD: University of Maryland. (DTIC No. AD A061569)

Stewart, A. L., Ware, J. E., Jr., & Brook, R. H. (1977, March). A study of the reliability, validity, and precision of scales to measure chronic functional limitations due to poor health. Santa Monica, CA: The Rand Corporation. (DTIC No. AD A043261)

Stinson, J., & Stokes, J. (1980, June). How to multi-appraise. Management Today, 43-53.

Stouffer, S. A. (1955). Communism, conformity, and civil liberties. Garden City, NY: Doubleday.
Wording of Items and Tone of Wording

Strahan, R. F. (1980, October). More on averaging judges' ratings: Determining the most reliable composite. Journal of Consulting and Clinical Psychology, 48(5), 587-589.

Strahan, R. G. (1982). Assessing magnitude of effect from rank-order correlation coefficients. Educational and Psychological Measurement, 42(3), 763-765.

Strang, H. R. (1980). Effect of technically worded options on multiple-choice test performance. Journal of Educational Research, 73(5), 262-265.
Wording of Items and Tone of Wording

Straton, R. G., & Catts, R. M. (1980). A comparison of two, three and four-choice item tests given a fixed total number of choices. Educational and Psychological Measurement, 40, 357-365.

Subkoviak, M. J., & Roecks, A. L. (1976, Winter). A closer look at the accuracy of alternative data-collection methods for multidimensional scaling. Journal of Educational Measurement, 13(4), 309-317.

Swan, J. E., Epley, D. E. (1981, February). Completion and response rates for different forms of income questions in a mail survey. Perceptual and Motor Skills, 52(1), 219-222.

Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. Journal of Experimental Psychology, 7, 456-461.
Number of Scale Points

Sympson, J. B., Weiss, D. J., & Ree, M. J. (1982, August). Predictive validity of conventional and adaptive tests in an Air Force training environment (AFHRL-TR-81-40). Minneapolis, MN: Department of Psychology, University of Minnesota. (DTIC No. AD A119031)

Takane, Y. (1981, March). Multidimensional successive categories scaling: A maximum likelihood method. Psychometrika, 46(1), 9-28.

Tate, P. (1984). The blossoming of European AI. Datamation, 30(18), 85-88.
Future Research

Tatsuoka, K. (1979, July). Analytical test theory model for time and score (CERL Report E-8). Urbana, IL: Computer-Based Education Research Laboratory, University of Illinois. (DTIC No. AD A076601)

Tatsuoka, K. K., Birenbaum, M., Tatsuoka, M. M., & Baillie, R. (1981, August). Psychometric approach to error analysis on response patterns of achievement tests. Catalog of Selected Documents in Psychology, 11(58).

Tatsuoka, K. K., & Tatsuoka, M. M. (1982, Fall). Detection of aberrant response patterns and their effect on dimensionality. Journal of Educational Statistics, 7(3), 215-231.

Teel, K. S. (1978, July). Self-appraisal revisited. Personnel Journal, 364-367.

Thompson, B. (1980). Comparison of two strategies for collecting Q-Sort data. Psychological Reports, 47, 547-551.

Thompson, J. A., & Wilson, S. L. (1982, October). Automated psychological testing. International Journal of Man-Machine Studies, 17(3), 279-289.

Thorton, G. C. (1980). Psychometric properties of self-appraisals of job performance. Personnel Psychology, 33, 263-271.

Thurstone, L. L. (1927). A law of comparative judgment. Psychological Review, 34, 273-286.

Thurstone, L. L. (1929). Fechner's law and the method of equal-appearing intervals. Journal of Experimental Psychology, 12, 214-224.

Towsend, J. W. (1979). A Guttman analysis of scales developed by retranslation. Dissertation Abstracts International, 39(11-B), 5627-5628.

Trollip, S. R., & Anderson, R. I. (1982). An adaptive private pilot certification exam. Aviation, Space, and Environmental Medicine, 53(10), 992-995.
Future Research

Turney, J. R., & Cohen, S. L. (1976, June). The development of a work environment questionnaire for the identification of organizational problem areas in specific Army work settings (ARI Technical Paper-275). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (DTIC No. AD A028241)

Uhlaner, J. E., & Drucker, A. J. (1980). Military research on performance criteria: A change of emphasis. Human Factors, 22(2), 131-139.

Urry, V. W. (1977, August). Tailored testing: A spectacular success for latent trait theory (TS-77-2). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center. (NTIS No. PB 274 576)

Using behaviorally anchored rating scales (BARS) (1980). Small Business Report, 16-19.

Vale, C. D. (1981, August). Design annd implementation of a microcomputer-based adaptive testing system. Behavior Research Methods and Instrumentation, 13(4), 399-406.

Vance, R. J., Kuhnert, K. W., & Farr, J. L. (1978). Interview judgments: Using external criteria to compare bahavioral and graphic scale ratings. Organizational Behavior and Human Performance, 22, 279-294.
Interviewing

Van Heerden, J., & Hoogstraten, J. (1980, February). Response preference as a function of instructions in an unstructured questionnaire. Perceptual and Motor Skills, 50(1), 227-230.

van Rijn, P. (1980, June). Self-assessment for personnel examining: An overview (Personnel Research Report 80-14). Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development Center, Alternatives Task Force.
Ethnic Background

Verity, J. W. (1984). AI tools arrive in force. Datamation, 30(15), 44-53.
Future Research

Vidali, J. J. (1976). Reliability of the semantic differential under practical conditions. Psychological Reports, 39, 583-586.
Bipolar Scales
Semantic Differential Scales

Volans, P. J. (1982, October). Pros and cons of tailored testing: An examination of issues highlighted by experience with an automated testing system. International Journal of Man-Machine Studies, 17(3), 301-304.

Walizer, D. G., & Mietus, J. R. (1980, March). Development of an organizational survey feedback program for the 32D Air Defense Command (ARI-TR-433). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences. (AD A100 972)

Warm, T. A. (1978, December). A primer of item response theory (Technical Report 941278). Oklahoma City, OK: U.S. Coast Guard Institute. (DTIC No. AD A063072)
Future Research

Warmke, D. L., & Billings, R. S. (1979). A comparison of training methods for altering the psychometric properties of experimental and administrative performance ratios. Journal of Applied Psychology, 64, 124-131.

Weeks, M. F., & Moore, R. P. (1981). Ethnicity-of-interviewer effects on ethnic respondents. Public Opinion Quarterly, 45, 245-249.
Ethnic Background
Interviewing

Weiss, D. J. (1982, Fall). Improving measurement quality and efficiency with adaptive theory. Applied Psychological Measurement, 6(4), 473-492.

Welch, S., Comer, J., & Steinman, M. (1973). Interviewing in a Mexican-American community: An investigation of some potential sources of response bias. Public Opinion Quarterly, 37, 115-126.
Ethnic Background
Interviewing

Welsh, J. R. (1977, February). An investigation into the sources of halo error. Dissertation Abstracts International, 37(8-B), 4203.

Wherry, R. J., Sr., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. Personnel Psychology, 35, 521-551.

White-Blackburn, G. Blackburn, I. C., & Lutzker, J. R. (1980). The effects of objective versus subjective quiz items in a Psi course. Teaching Psychology, 7(3), 150-153.

Whitely, S. E. (1977). Relationships in analogy items: A semantic component of a psychometric task. Educational and Psychological Measurement, 37, 725-739.

Wiegand, D. (1983). Present standardization of test procedures for ergonomic test of tracked and wheeled vehicles of the Federal Armed Forces Germany. Federal Armed Forces Proving Ground 41. Unpublished paper.

Wienclaw, R. A., & Hines, F. E. (1982, November). A model for determining cost and training effectiveness tradeoffs. Training Equipment Interservice/Industry Training Equipmment Conference, 405-416.
Multiple-Choice Scales

Wilcox, R. R. (1982, Summer). Bounds on the k out of n reliability of a test, and an exact test for hierarchically related items. Applied Psychological Measurement, 6(3), 327-336.

Wilcox, R. R. (1982). Determining the length of multiple choice criterion-referenced tests when an answer-until-correct scoring procedures is used. Educational and Psychological Measurement, 42, 789-794.

Williams, J. A. (1964). Interviewer-respondent interaction: A study of bias in the information interview. Sociometry, 27, 338-352.
Interviewing

Wind, Y., & Lerner, D. (1979). On the measurement of purchase data: Surveys versus purchase diaries. Journal of Marketing Research, 16, 39-47.

Winkler, J. D., Kanouse, D. E., & Ware, J. E., Jr. (1982). Controlling for acquiescence response set in scale development. Journal of Applied Psychology, 67(5), 555-561.

Wise, S. L. (1982). A modified order-analysis procedure for determining unidimensional item sets. Dissertation Abstracts International, 42(7-A), 3121.

Wood, J. A. (1982). The quantification of verbal anchors used to denote occurrences of frequency, amount, and evaluation on five response category Likert scales. Dissertation Abstracts International, 43(2-A), 408-409.

Yadav, M. S., Govinda, R., & Thomas, K. T. (1976). Some psychometric studies in attitude scale construction. Psychological Studies, 21(1), 1-11.

Young, F. W., & Levinsoh, J. R. (1974). Two special-purpose programs that perform nonmetric multidimensional scaling. Behavior Research Methods and Instrumentation, 6(3), 354-355.

Zammuto, R. F., London, M., & Rowland, K. M. (1982). Organization and rater differences in performance appraisals. Personnel Psychology, 35, 643-658.

Zedeck, S. (1981). Behaviorally based performance appraisals. Aging and Work, 4(2), 89-100.

Zedeck, S., & Cascio, W. F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. Journal of Applied Psychology, 67(6), 752-758.

Zedeck, S., Jacobs, R., & Kafry, D. (1976). Behavioral expectations: Development of parallel forms and analysis of scale assumptions. Journal of Applied Psychology, 61, 112-115.

Zedeck, S., & Kafry, D. (1977, April). Capturing rater policies for processing evaluation data. Organizational Behavior and Human Performance, 18(2), 269-294.

Zedeck, S., Kafry, D., & Jacobs, R. (1976). Format and scoring variations in behavioral expectation evaluations. Organizational Behavior and Human Performance, 17, 171-184.

Multiple-Choice Scales
Questionnaire Layout

Appendix A

P-77-2

Questionnaire Construction Manual Annex
Literature Survey and Bibliography

Table of Contents

Not every topic covered in P-77-2 is covered in this sequel. Appendix A provides the reader with a way to reference back to the original work of P-77-2, Questionnaire Construction Manual Annex. This may be useful for situations where researchers prefer to compare earlier questionnaire survey research with the more current literature.

TABLE OF CONTENTS

Chapter		Page
I	INTRODUCTION	I-1
II	ADVANTAGES AND DISADVANTAGES OF VARIOUS TYPES OF QUESTIONNAIRES	II-1
	Methods to Measure Attributes and Behavior	II-1
	Comparison of the Structured Interview and Mail Questionnaires	II-1
	Comparison of the Structured Interview and Other Questionnaires	II-3
	Comparison of Open- and Closed-Ended Items	II-4
	Conclusions	II-5
III	SELECTION OF QUESTIONNAIRE ITEMS	III-1
	Content of Questionnaire Items	III-1
	Methods for Determining Questionnaire Content	III-1
	Other Considerations Related to Questionnaire Content	III-2
	Pros and Cons of Various Types of Questionnaire Items	III-3
	Ranking Items	III-3
	Rating Scale Items	III-5
	Multiple Choice Items	III-9
	Forced Choice and Paired Comparison Items	III-11
	Card Sorts	III-14
	Semantic Differential Items	III-15
	Other Types of Items	III-17
	Conclusions Regarding the Pros and Cons of Various Types of Questionnaire Items	III-19
IV	COMPARISON OF SCALING TECHNIQUES	IV-1
V	EFFECTS OF VARIATION IN PRESENTATION OF QUESTIONNAIRE ITEMS	V-1
	Mode of Items	V-1
	Wording of Items	V-1
	Clarity of Items	V-13
	Difficulty of Items	V-15
	Length of Question Stem	V-18
	Order of Question Stems	V-19
	Order of Response Alternatives	V-25
VI	NUMBER OF RESPONSE ALTERNATIVES AND RESPONSE ANCHORING	VI-1
	Issues Regarding Number of Response Alternatives to Employ	VI-1
	Response Anchoring	VI-9

TABLE OF CONTENTS (Cont.)

Chapter		Page
VII	ORDER OF PERCEIVED FAVORABLENESS OF COMMONLY USED WORDS AND PHRASES	VII-1
	Major Studies and Lists of Adjectives and Scale Values	VII-1
	Summary and Conclusions	VII-29
VIII	CONSIDERATIONS RELATED TO THE PHYSICAL CHARACTERISTICS OF QUESTIONNAIRES	VIII-1
	Location of Response Alternatives Relative to Stem	VIII-1
	Questionnaire Length	VIII-1
	Questionnaire Format Considerations	VIII-2
	The Use of Answer Sheets	VIII-3
IX	CONSIDERATIONS RELATED TO THE ADMINISTRATION OF QUESTIONNAIRES	IX-1
	Effects of Instructions	IX-1
	Effects of Various Motivational Factors	IX-2
	Effects of Anonymity	IX-6
	Effects of Administration Time	IX-9
	Effects of Characteristics of Questionnaire Administrators	IX-10
	Effects of Administration Conditions	IX-13
	Effects of Other Factors Related to Questionnaire Administration	IX-14
X	CHARACTERISTICS OF RESPONDENTS THAT INFLUENCE QUESTIONNAIRE RESULTS	X-1
	Item Format Biases	X-1
	Social Desirability Response Set	X-2
	Acquiescence Response Set	X-3
	Extreme Response Set	X-4
	Effects of Attitudes on Responses	X-5
	Effects of Demographic Characteristics on Responses	X-6
	Summary and Conclusions	X-7
XI	CONSIDERATIONS RELATED TO THE EVALUATION OF QUESTIONNAIRE RESULTS	XI-1
	Scoring of Questionnaire Results	XI-1
	Properties and Uses of Ipsative Scores	XI-3
	Data Analyses	XI-6

TABLE OF CONTENTS (Cont.)

Chapter		Page
XII	RECOMMENDED AREAS FOR FURTHER RESEARCH	XII-1
	Advantages and Disadvantages of Various Types of Questionnaires	XII-1
	Selection of Questionnaire Items to be Used	XII-1
	Comparison of Scaling Techniques	XII-2
	Effects of Variation in Presentation of Questionnaire Items	XII-2
	Number of Response Alternatives and Response Anchoring	XII-3
	Order of Perceived Variables of Commonly Used Words and Phrases	XII-3
	Considerations Related to the Physical Characteristics of Questionnaires	XII-3
	Considerations Related to the Administration of Questionnaires	XII-3
	Characteristics of Respondents that Influence Questionnaire Results	XII-4
	Considerations Related to the Evaluation of Questionnaire Results	XII-4
	General Recommendations	XII-4
	BIBLIOGRAPHY	B-1

Appendix B

Comparison Between P-77-2, Questionnaire Contruction Manual Annex, and the Sequel

This appendix delineates the content areas covered in P-77-2, and in this sequel. Each content area is identified by where it can be found in the sequel by title of the section, and then by what chapters it can be found in P-77-2. The usual case has been that a stand-alone section in the sequel can be found in more than one chapter of P-77-2. Some content areas are included in the sequel, but were not part of P-77-2. In addition, there are other content areas that were covered in P-77-2, but were not included in the sequel.

In P-77-2, common scaling techniques were found in two different chapters. Chapter III, Selection of Questionnaire Items to Be Used, compared various types of questionnaire items such as: ranking and rating scale items, paired comparisons, card sorts, semantic differential, checklists, multiple choice, and forced choice. Chapter IV compared the above mentioned scaling techniques. The sequel compares and updates the research for some of the same scaling techniques: multiple choice (this section includes Likert scales, Guttman scales, checklist, Q Sort, and behavioral scales), bipolar, semantic differential, rank order, and paired comparison.

A new addition to the literature has been included for the behavioral scales. The foundation for behavioral scales is the critical incident. The critical incident technique is mentioned in Chapter III of P-77-2. However, Behaviorally Anchored Rating Scales are not discussed in P-77-2. Behaviorally Anchored Rating Scales have been expanded so that there are now a wide variety of methods for this type of scale development and numerous forms of behavioral scales. The sequel includes sections on Behaviorally Anchored Rating Scales, Behavioral Expectation Scales, Behavioral Observation Scales, and Mixed Standard Scales. These behavioral scales were originally developed to encourage raters to observe behavior more accurately. They have been primarily used in questionnaire construction for performance appraisal purposes. Even so, there has been research which indicates that they have a broader application that includes surveys.

Format differences were discussed in P-77-2, Chapter III, Selection of Questionnaire Items, as to the pros and cons of various types of questionnaire items, in Chapter VIII, Considerations Related to the Physical Characteristics of Questionnaires, and in Chapter X, Characteristics of Respondents that Influence Questionnaire Results. In the sequel, questionnaire format has been addressed in Chapter VII as a stand-alone chapter.

A format difference for questionnaires has been included in the sequel in Section 7.2, Branching. Branching is a common approach used by researchers to guide respondents through a questionnaire to some questions, but not necessarily to all questions. Branching is also synonymous with other terms such as leading and routing. This topic area was not included in P-77-2.

P-77-2 incorporates the number of response alternatives to use in questionnaires, and the response anchoring, in Chapter VI, Number of Response Alternatives and Response Anchoring. The sequel separates out these two topic areas into three independent sections: Section 5.3, Number of Scale Points; Section 5.1, Response Alternatives; and Section 5.4, Middle Scale Point Position. The middle scale point position was incorporated into P-77-2 in Chapter VII, Order of Perceived Favorableness of Commonly Used Words and Phrases, and also in Chapter VI, Number of Response Alternatives and Response Anchoring.

The sequel includes a stand-alone section, Section 2.6, Continuous and Circular Scales, which was not part of P-77-2. Continuous scales have no scale points. The rationale for their use is that they will yield greater discrimination by raters. Circular scales are scales that were structured in a circumplex, or circle, to eliminate errors of extreme judgments, and errors of central tendency.

Open- and closed-end items are discussed in P-77-2 in Chapter II, Advantages and Disadvantages of Various Types of Questionnaires, and in Chapter III, Selection of Questionnaire Items to be Used. In the sequel, this topic area has been expanded and is found in Section 4.1, Open-Ended Items and Closed-End Items. Balancing response alternatives and the positive and negative wording of items is included in P-77-2 in Chapter V, Effects of Variation in Presentation of Questionnaire Items, in Chapter VI, Number of Response Alternatives and Response Anchoring, and in Chapter X, Characteristics of Respondents that Influence Questionnaire Results. These content areas are combined in the sequel to Chapter 4.5, Balanced Items.

The length of a questionnaire, the number of items in a questionnaire, and the number of words in an item are included in the sequel in Section 4.3, Length of Items and Number of Items. This content area is covered in P-77-2 in Chapter V, Effects of Variation in Presentation of Questionnaire Items, and in Chapter VIII, Considerations Related to the Physical Characteristics of Questionnaires. The ordering of items is found in P-77-2 in Chapter V, Effects of Variation in Presentation of Questionnaire Items, and in the sequel in Section 4.4, Order of Items. The sequel includes Section 4.2, Wording of Items and Tone of Wording. This content area may be found in P-77-2, Chapter V, Effects of Variation in Presentation of Questionnaire Items, and Chapter VII, Order of Perceived Favorableness of Commonly Used Words and Phrases.

Interviewing is treated as an independent topic in the sequel in Section 6.1. In P-77-2, interviewing is discussed in Chapter II, Advantages and Disadvantages of Various Types of Questionnaires, and in Chapter IX, Considerations Related to the Administration of Questionnaires.

In recent years, there has been a trend away from research that focuses on questionnaire construction for content areas such as: format, number of scale points, and types of response alternatives. A greater focus has been placed on respondent demographic characteristics that might influence a rating, training respondents in how to rate, the complexity of the questionnaire, and the cognitive complexity of the rater. The theory of cognitive complexity has its foundation in the work of Kelly (1955), and has been defined as the ability to differentiate person-objects in the

social environment. The sequel includes Section 6.2, Cognitive Complexity. This topic is not covered in P-77-2.

Demographic characteristics that describe respondents in questionnaire construction have been divided into four sections in the sequel, and these sections are: Section 6.3, Education; Section 6.4, Ethnic Background; Section 6.5, Gender; and Section 6.6, Age. These demographic characteristics are found in P-77-2 in Chapter IX, Considerations Related to the Administration of Questionnaires, and In Chapter X, Characteristics of Respondents that Influence Questionnaire Results.

P-77-2 and the sequel both cover the same content areas in most instances. There are some areas where there is not overlap. For example, P-77-2 does not include material on Behaviorally Anchored Rating Scales, Continuous and Circular Scales, and Cognitive Complexity, while the sequel does. The sequel does not include areas on Questionnaire Administration and Evaluation of Questionnaire Results, which are included in P-77-2. (See Appendix C for an overview of content areas covered by P-77-2 and the sequel.)

Appendix C

Overview of Content Areas Covered by P-77-2 and the Sequel

Questionnaire Construction Content Areas	P-77-2	Sequel
Scaling Techniques	Yes	Yes
Behaviorally Anchored Rating Scales	No	Yes
Format	Yes	Yes
Branching	No	Yes
Response Alternatives	Yes	Yes
Continuous and Circular Scales	No	Yes
Open- and Closed-End Items	Yes	Yes
Balancing Response Alternatives	Yes	Yes
Wording of Items	Yes	Yes
Length and Number of Items	Yes	Yes
Order of Items	Yes	Yes
Interviewing	Yes	Yes
Cognitive Complexity	No	Yes
Demographic Characteristics	Yes	Yes
Administration of Questionnaires	Yes	No
Evaluation of Questionnaire Results	Yes	No
Further Research	Yes	Yes

Appendix D

Future Research Recommendations

Future research recommendations have been grouped together according to content area by chapter: Chapter II, Scale Categories; Chapter III, Behavioral Scales; Chapter IV, Design of Questionnaire Items; Chapter V, Design of Scale Categories; Chapter VI, Interviewer and Respondent Characteristics; and Chapter VII, Questionnaire Format. These groupings are not ranked. The intention is to provide an overview of each content area for what types of research might be performed. These recommendations were selected because of perceived gaps in survey research for specific content areas.

Chapter II. Scale Categories

- Practical, workable procedures are sorely needed to focus investigators on refining and implementing developmental procedures in the construction of questionnaires, instead of arbitrarily borrowing items from other surveys. Development and validation of such procedures should be addressed.

- No one scale category can be recommended over another. More research is required to replicate studies in order to establish which is the best scale category to use for specific types of applications.

- The combination of scales, such as ordinal and interval scales, used in conjoint measurement requires further refinement.

- Guttman scaling theory requires further exploration as it applies to developing interviews which are predictive of future behavior, physical health, etc. These scales are probably the most difficult to develop of all the scaling techniques covered in this report.

- Research is needed to find ways to develop Guttman scales using fewer subjects and in shorter time frames. This would be a major breakthrough for expanding the use of Guttman scales.

- More studies are needed to determine whether subjects are confounding trait dimensions with response alternatives.

- Replication of studies is required to determine whether subjects are influenced by descriptive anchors, thereby making greater use of the categories at the extreme ends of the scale.

- Further evidence is required to substantiate the existence of response style as it relates to the order of item presentation.

- It cannot be concluded that adverb and numerical scales measure meaning in the same way. Factor analysis would allow for a better understanding of the meaning of the items. Additional studies to confirm/disconfirm the findings would be of value.

- More research will be required to determine reliable and valid procedures for transforming ordinal data into interval scales in conjoint measurement.

- Data collection for value profiles has more or less abandoned the rank order technique for the more popular Likert scaling method. Yet, there is evidence to support the use of rank ordering or paired comparison methods as being more reliable. Test-retest reliability for the scales mentioned above needs further investigation by comparing rho and Kendall's tau in the statistical analysis. In addition, software could be developed which would make this technique easier and more quickly performed.

- Future studies which compare rank order methods with other scale methods should consider experimental designs with repeated measures since an individual's preference for type of scale may have affected the stability of the results over time. Experimental designs should focus on the effects of individuals, the effects of method, and the individual method interaction effects.

- Researchers have extended the use of card sorts for marketing purposes. It was found that subjects were able to reconstruct the underlying situational hierarchical structures on a card sort for the first trial. Second trial card sorts shifted to a less ordered structure. Further investigation will be required to resolve why this type of shift was observed.

- Continuous scales have been used to provide greater discrimination by respondents. However, more research is needed since there has been evidence that respondents may be rating continuous scales on the equivalent of five or six categories depending on what is being measured.

- Continuous scales have been integrated and transformed into category scales in the measurement of psychophysical phenomena. The combination of these two scales requires more research for measuring vehicle environments, such as vibration or temperature.

Chapter III. Behavioral Scales

- The application of Behaviorally Anchored Rating Scales (BARS) to measure group morale was encouraging as an alternative to the traditional self-report measures obtained in surveys. Other applications for BARS surveys may be useful in reducing the subjectivity found in other self report instruments.

- When BARS have been used solely for performance appraisal, the time and cost factors involved in their development have been high. Research is needed to identify multiple uses for BARS, such as delineating organizational goals, in order to make development efforts worthwhile.

- Further research is needed to validate rater training programs using BARS to increase rater accuracy.

- Further research is needed to investigate the possible increase in validity for BARS when multiple rater groups stipulate the dimensions of performance that effects them.

- Most empirical research dealing with performance evaluation systems, such as Behavioral Expectation Scales (BES), measures improvements in attitudes and performance for short study periods of less than one year. Longitudinal studies for the impact of BES over time would be worthwhile in establishing or disputing the long-range results that these systems hope to have.

- BES require further investigation as to behavior change, relationship to performance effectiveness, and developmental and scoring procedures.

- More studies are needed to test the effects of various training interventions on psychometric error, and the use of BES.

- Behavioral Observation Scales require more research to determine whether these scales are really measuring simple observations or whether they are measuring trait-like judgment due to rater recall over time.

- Mixed Standard Scales (MSS) research should focus on the rating process and environment since the psychometric properties of the rating scales may be influenced by organizational reward practices and organizational climate.

- MSS have been shown to be reliable and valid. However, raters prefer other scale types over MSS. Additional research is needed to determine whether MSS are appropriate for use in operational test and evaluation (OT&E).

- Anchor items on MSS require investigation for possible multidimensionality since this would yield inconsistent ratings.

Chapter IV. Design of Questionnaire Items

- Further investigation is needed to substantiate whether respondents with more education will answer open-ended and closed-end forms consistently. Respondents with less education do not respond consistently to these form differences.

- There has not been consistent replication as to the wording of items and the tone of wording so that researchers have not been able to predict when the items will be influenced by the wording and which words will influence the results. More research is needed in this area.

- Researchers need to assess various procedures or methods as a way to identify the use of specific words which are used in items. Such a method may have the potential to identify the structure of the item itself.

- Research is needed in the OT&E community to determine whether item reduction techniques used in marketing surveys and behavioral scale development would be applicable in reducing the number of items used in field surveys.

- Methodological research for understanding why order effects occur is an important area for questionnaire construction research to examine. For example, it is assumed that separating two items by several or many others may eliminate a known context effect. Further investigation and greater understanding of these issues is needed.

- More needs to be known about the usefulness of balancing anchors in conjunction with what type of application the scale will have.

- Replication would be helpful to indicate whether items of greater length will elicit responses toward the middle of the scales. In addition, replication is required to determine whether items of short length elicit responses toward the positive end of the scale.

- Further replication is suggested for long items which are negatively worded to determine whether they elicit responses toward the midrange of the scale.

- More studies would be desirable to investigate both positive and negative items on a questionnaire. There has been some evidence that negatively worded items may result in less accurate responses, and reduce response validity.

Chapter V. Design of Scale Categories

- There has been some evidence that formats which are easiest to rate would be best for respondents who have lower levels of education. Further research is needed to explore this finding, and to identify which formats those might be.

- If response alternatives are selected independent of the item, and measured for bands along the scale dimension, there may be the possibility that response alternative linkage to the item may modify the standard deviation of each response. The linkage of the response alternative to the item for variations in the standard deviation needs further investigation.

- More studies would be useful in examining whether fully labeled scales yield less skewed responses than scales anchored only at the extremes. There has been some evidence that fully labeled scales achieved higher test-retest correlations. This finding was observed with market-segmentation studies, and would need investigation for generalizing to OT&E studies.

- There has been contradictory evidence that verbal anchors may lower accuracy in psychophysical task ratings. More thorough and systematic research would be useful in establishing the effects of verbal anchoring on psychometric criteria for psychophysical tasks.

- Pictorial anchors have been subjected to limited investigation. This methodology needs to be extended to different types of visually perceived stimuli in OT&E.

- There has been conflicting evidence regarding individuals who may or may not respond to the "Don't Know" category. One theory has proposed that there is a unique constellation of traits which describes individuals who "float" back and forth between the "Don't Know" response, and other response alternatives. "Floating" may not have a single cause, although there is no model to describe subjects who exhibit this behavior. Additional investigation is needed.

- Further studies are needed to identify the respondent characteristics and/or the processes involved in responses to the presence or absence of a middle alternative.

Chapter VI. Interviewer and Respondent Characteristics

- More studies would be useful to improve interview techniques which are used in telephone surveys. Experimental variations of question presentation may be useful.

- Military survey research that incorporates interviews needs to investigate ways to obtain more lead time in survey development, increase response rate for large periodic surveys, control standardization in field administration, and control for response bias of personnel brought about by the influence of superiors.

- More research is needed on the effects of third parties on face-to-face interviews.

- To investigate cognitive complexity, the contextual differences for the type of organization, and the characteristics of the respondents, might account for the failure to replicate previous results. Organizations other than institutions of high education may be a more appropriate environment to investigate whether cognitive complexity is a relevant variable in the rating task. Cognitive complexity issues should be investigated for OT&E respondent populations.

- Further replication is needed on the generalizability of item rating by level of education from the general U.S. population to military personnel.

- There may be an interaction among educational level of respondents, response consistency in rating items, and intensity of feeling about items. Investigation of this interaction is needed to replicate previous research.

- Educational level and item nonresponse may be related; this area could use more research. This type of survey is difficult to design since it is challenging to obtain strategies for identifying nonrespondents.

- Very little is known about the effect of education on questionnaires that are designed for use in performance appraisal. Research in this area may be useful.

- There has not been consistency in findings for the impact of education and questionnaire construction. More research is needed on the influence of combined relationships with other demographic variables, as well as generalizability of these results to military personnel.

- Results have been mixed as to ethnic background of raters assessing ratees on performance appraisal scales. Further research is needed for the effect of same-race raters and different-race raters.

- Further research may be required to determine the extent to which self assessment can be used, and the influence of ethnic background on self assessment.

- Not enough is known regarding the impact of ethnic background on questionnaire construction. More studies for the influence of ethnic background and questionnaire construction would be useful in designing and analyzing these instruments.

- Review of the research relating to questionnaire construction and the differentiation of response patterns by gender of respondents has received mixed results. Some studies have found differences in rating by females and males, while other studies have not. Research is needed to identify interactions of gender, education, age, and ethnic background. For example, investigation of item content and gender may be useful since the content of an item may have the potential to bias survey results where males and females hold different values.

- For opinion surveys, further studies would be useful in examining the influence of age on item content to assess the historical perspective of the different groups.

- Nonresponses to items following a branch, and survey nonresponse, may be influenced by age and education of the respondent. Replication would be beneficial in assessing this phenomenon as it is influenced by demographic characteristics of Army populations in OT&E.

Chapter VII. Questionnaire Format

- Evidence supporting any one format was sparse and inconsistent. More studies may be useful in identifying strengths and weaknesses of different scale formats. However, other methodological issues may have greater potential for research such as: Rigor of developmental procedures, preference of format by raters, matching characteristics of rater and format, assessment of scoring systems, and examination of different sampling techniques.

- A major implication for future research appears to be the reassessment of scale item selection. This could be approached in two ways. First, through a technique that would allow for the wording of each item (e.g., the Echo method). Second, item reduction procedures that maintain contruct validity.