Research, Consultation, & Teaching Program Monograph No. 6

Construction and Use of Behavior Rating Scales

Changnam Lee
Gerald Tindal

Behavioral Research and Teaching, College of Education
5262 University of Oregon
Eugene, Oregon, 97403-5262

Published by
Research, Consultation, and Teaching Program
Behavioral Research and Teaching
College of Education
University of Oregon

Staff
Gerald Tindal, Program Director
Jerry Marr, Editor
Abe Deffenbaugh, Assistant Editor

Lee, Changnam; Tindal, Gerald
Construction and Use of Behavior Rating Scales
Monograph 6

Preparation of this document was supported in part by the U.S. Department of Education, grant number H029B20030. Opinions expressed herein do not necessarily reflect the position or policy of the U.S. Department of Education, and no official endorsement by the Department should be inferred.

Layout: Jerry Marr
Cover design: George Beltran
Editing assistance: Gretchen Matis

Construction and Use of Behavior Rating Scales

Changnam Lee
Gerald Tindal

University of Oregon

Abstract

This monograph presents a comprehensive review of behavioral rating scales from four vantages. Scale construction is considered with three major topics addressed: design, particularly content and format; development, with descriptions given for selecting appropriate statements to include in a scale and ensuring their proper placement and contribution to the total scale outcomes; and finally, evaluation, including use of quantitative, qualitative, and use criteria. The last half of the paper presents an analysis of several popular rating scales that are commercially available; two perspectives are framed: self and informant. By tying these four topics together, this monograph presents the premise that, not only must professionals be skilled in adoption of appropriate behavioral rating scales, but more importantly, they also must be knowledgeable about how rating scales are constructed and how results can be interpreted.

Introduction

Because of the importance of behavior assessment in research and behavior modification programs, and given the demand for systematic approaches to it, many assessment techniques and procedures have been developed, such as direct observation, self- monitoring, behavioral interviews, and assessment with rating scales. Each of these approaches has unique advantages over others, as well as shortcomings. In this essay, we will focus on the assessment with rating scales.

As Edelbrock (1988) suggested, the rating-scale method has some unique advantages. First, behavioral checklists, or rating scales, provide standardized descriptors of a specific behavior. These descriptors can be used as critical factors in designing programs for behavior modification, in classification and training, and in direct observation. Secondly, they provide a means for rating the presence, frequency, and severity of specific behaviors, and for rating global attributes. Third, they are simple, fast, and economical in terms of both cost and professional time. Therefore, they make it possible to assess a large group of people for epidemiological studies, screening, and program evaluations. Fourth, they provide quantitative indexes of child functioning at the baseline or in response to interventions, and they can be a means of determining whether a child's behavior is appropriate or deviant in relation to normative groups.

In using behavioral assessment instruments such as checklists and rating scales, we need to make certain assumptions (Edelbrock, 1988). The first assumption is that the rater or informant understands the behavioral construct being rated and knows which behaviors are concerned with that construct. If a rater has a vague or ambiguous idea about the construct or domain, the instrument is likely to yield unreliable and invalid information. Therefore, either in the development or in the implementation of a rating instrument, the construct should be clearly defined.

The second assumption is that universal agreement has been achieved regarding the reference points for scaling those behavioral ratings. A rating scale is an attempt to quantify an attribute; therefore a systematic scoring procedure must be provided to reduce errors in psychometric properties, as will be discussed below.

The third assumption in using rating scales is that

the rater or informant is able to "extract a cumulative impression of the target construct from the stream of everyday life activities" (p. 352). Therefore, prior to rating any behaviors, we should make a considerable observation of the individual whose behaviors are to be rated.

Many scales have been developed and used for a wide variety of purposes. While some researchers created their scales using systematic procedures, many others have developed theirs in an arbitrary fashion (Rie & Friedman, 1978). As a result, rating scales may not have adequate reliability or validity. In addition, most scales are designed to assess people of a specific age range, or specific characteristics, and are inappropriate for other groups of people. A new rating instrument must then be developed. Therefore, systematic procedures should be provided to construct both reliable and valid instruments.

In addition to construction of instruments, selecting an instrument is another important issue. Since many rating instruments have been published, we need to know how to select an appropriate one that serves a specific purpose. This monograph is designed to (a) provide future researchers with behavior rating scale construction procedures that have been proven to work, and (b) provide future users (especially teachers) of behavior rating scales some guidelines to select the most appropriate ones for their purposes.

Behavior Rating Scale Construction

As a term used in behavioral research, a scale can be defined as a collection of items, the responses to which are scored with numerical values and are combined to yield a summative score (adapted from Dawis, 1987). Therefore, a behavior rating scale (or a series of behavior rating scales) is an instrument that attempts to quantify behavioral dimensions or traits that individuals exhibit on the basis of some predefined behavior domain or construct.

Usually several items represent a behavior dimension. For example, the Child Behavior Rating Scale (Cassel, 1962) has five dimensions: (a) Self Adjustment, (b) Home Adjustment, (c) Social Adjustment, (d) School Adjustment, and (e) Physical Adjustment. Each of the first three dimensions has 20 items, the School Adjustment has 12 items, and the Physical Adjustment has 6 items. Each item has units and reference points of the units on a continuum. Such units as descriptive words, phrases, or statements are fixed on the behavioral continuum so that a rater can use them as criteria in deciding the relative magnitude of an individual trait. These established units with their reference points are usually called "anchors" (Torgerson, 1958, p. 79).

This section focuses on the construction of behavior rating scales. Various methods or approaches that have been developed and practiced are reviewed. Dawis (1987) divided the scale construction procedure into three stages: scale design, scale development, and scale evaluation. We incorporate these stages for our review.

Scale Design

Scale design includes content and format. We have to consider what type or aspect of behavior an item should describe, and how the items should be presented. Because a cluster of items typically represents a behavioral dimension, each item should describe a critical behavior so that all the items properly depict the dimension. In most cases, we can not easily determine what are the critical behaviors that constitute a behavioral dimension or trait. Therefore, scale content is necessarily related to the issue of the source of items and the method of collecting items. Scale format, on the other hand, structures the type of responses for an item. The format should be designed in such a way that respondents can easily understand what they should do, and errors in rating should be minimized.

Scale Content

The construct to be measured should be clearly defined with well-established procedures for sampling or selecting items or statements. Items could be collected from existing literature, previously developed measures, case records of disturbed children, and unstructured reports by teachers (Edelbrock, 1988). Open-ended interviews with representative subjects from the target respondent population can also be used (Dawis, 1987).

A unique and systematic interview method was introduced by Flanagan (1954), using the critical incident technique. It has been adapted by many performance rating scales, such as the Behaviorally Anchored Rating Scales (BARS), the Behavioral Observation Scales (BOS) and the Mixed Standard Scales (MSS), which are reviewed later in this chapter.

Critical Incident Technique

The critical incident technique grew out of studies in the Aviation Psychology Program of the United States Army Air Forces in World War II. Flanagan (1954) defined an *incident* as "any observable human activity that is sufficiently complete in itself to permit inferences and predictions to be made about the person performing the act" (p. 327), and *critical incidents* as incidents that could differentiate "success and failure in performing an important part of the job assigned in a significant number of instances" (p. 329) in defined settings. These two definitions assume that cliches and stereotypes exist in specific jobs and that similar incidents occur in similar jobs. Essentially this technique is a procedure for gathering certain critical facts pertaining to behavior in defined situations. In this procedure,

only "qualified" (p. 335) observers are included who are well acquainted with the job performance.

To gather valid items for a scale, general activity aims to be assessed are established by an authoritative observer. For example, respondents may be asked, "In a few words, how would you summarize the general aim of (a specified activity)?" (p. 337). After establishing such general aims, they are asked to observe the future ratees' job performance for a specified period of time, in specified situations. Then they are asked to extract critical incidents from the observation and classify them immediately to ensure accuracy.

Flanagan (1954) has suggested that the critical incident technique can be used as: (a) measures of typical performance on criteria, (b) measures of proficiency on standard samples, (c) training, (d) selection and classification (screening), (e) job design and purification, (f) operating procedures, (g) equipment design, (h) motivation and leadership, and (i) counseling and psychotherapy.

Scale Format

This second component of scale design is the configuration of response choices that are presented to the rater. Dawis (1987) suggested that rating response formats differ in two primary ways: (a) the number of scale points, and (b) the manner in which scale points are anchored. According to him, 2-, 3-, or 5-point scales are the most common. Generally, more scale points are better than fewer in terms of psychometric properties and can generate more variability. Excessive use of the middlemost scale point can be avoided by using an even number of scale points.

A variety of formats have been developed and used for anchoring scale points. Guilford (1954) has classified commonly used rating scales into five broad categories: (a) numerical, (b) graphic, (c) standard, (d) cumulated points, and (e) forced choice. Each of these is briefly described below.

Numerical Scales

The typical numerical scale provides the observer or rater with a sequence of "defined" numbers. Here the term "defined" indicates that each number is paired with a descriptive cue. In the course of rating, an appropriate number is assigned to each stimulus or statement in line with the definitions or descriptions. Figure 1 presents an example adapted from Guilford (1954, pp. 263-264).

How did you feel about the math classes?

1 – Very uncomfortable
2 – Moderately uncomfortable
3 – Indifferent
4 – Moderately enjoyable
5 – Very enjoyable

Figure 1. An example of numerical scale.

Some numerical scales do not provide overt numbers for the rater. In such cases the rater is asked to respond to the items by choosing the most appropriate descriptive cues, and the researcher assigns numbers to them. Figure 2 represents an example of this format.

Does the baby reach for familiar persons?

Never Seldom Sometimes Generally Always

Figure 2. A sample numerical scale item without overt numbers for response alternatives.

Numerical scales are sometimes called "Likert-type" formats (Dixon, Bobo & Stevick, 1984; Latham & Wexley, 1977), which may be either unidirectional or bidirectional (Aiken, 1985). In a unidirectional scale, one end represents a minimum amount and the other end a maximum amount of the behavior to be rated. In a bidirectional scale, two negative (positive) poles are provided, one at either end of the scale; the best (or the worst) amount of that variable is represented at the center of the scale. Figure 3 illustrates the difference between the two formats.

Frisbie and Brandenburg (1979) compared two pairs of scale types. In one pair of formats, the alternatives were lettered and those in the other pair were numbered; the alternatives in one format of each pair were defined and the other format was end-defined. The four formats are illustrated in Figure 4.

The group responding to the scale with only endpoints defined had higher mean ratings (more positive) on most of the items (6 out of 8) than the group responding to the scale with all points defined. This suggests that a fully defined scale is less liable to leniency errors, the tendency of rating individuals too high or too low. The authors concluded that, if only the endpoints of the response scale are defined, the items are not equivalent, and that, whether the response choices are numbered or lettered, the items are equal. They also concluded that scores from the two formats were not comparable with each other.

Dixon, et al. (1984), however, reported somewhat different results. They compared the effects of the "Likert-type defined format" and "Likert-type end-defined format" (p. 65). In this study, Scale I consisted of two parts: (Part A) 29 items in the Likert-type defined format, and (Part B) the same number of items in the Likert-type end-defined format. For Scale II, each

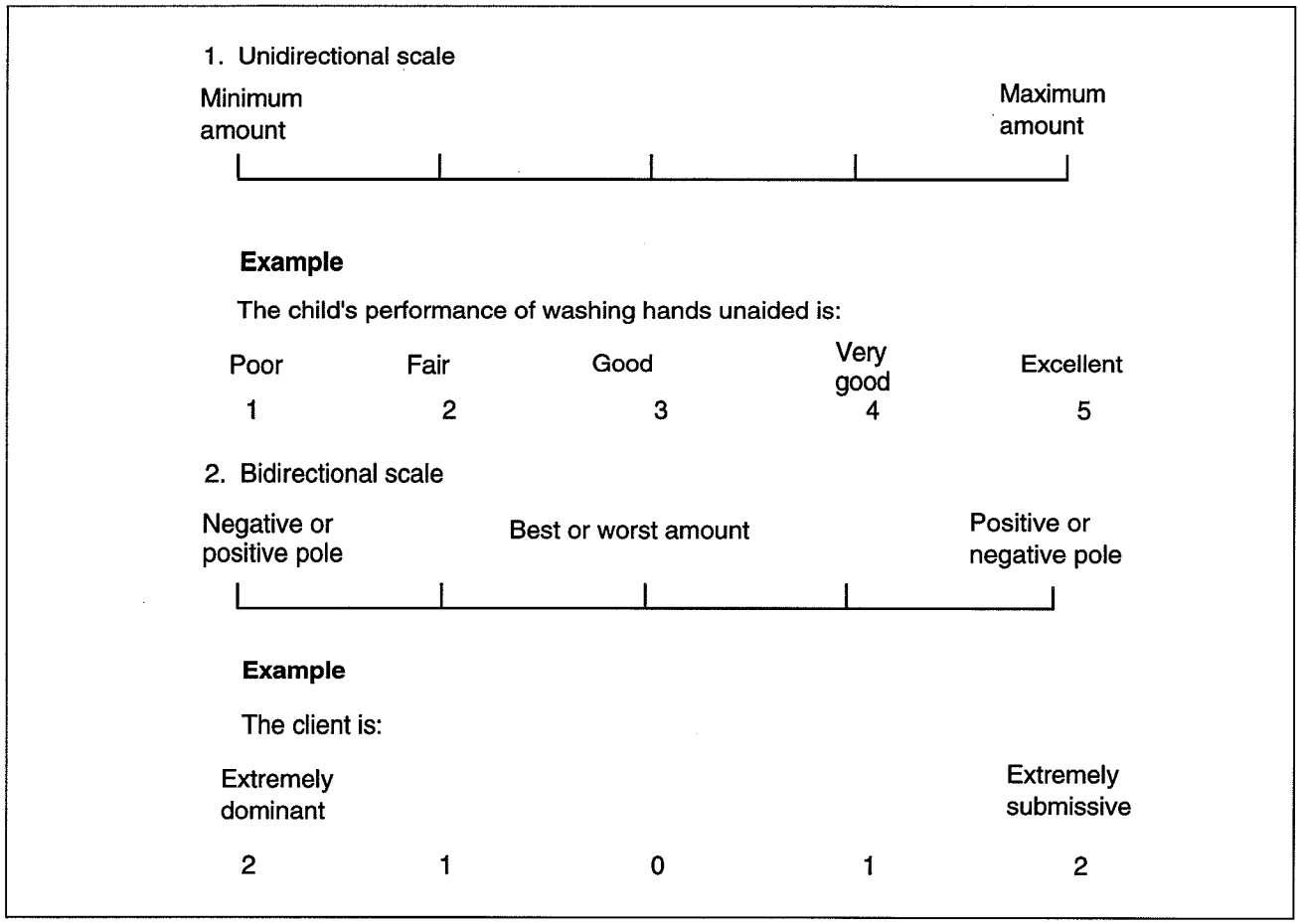

Figure 3. Illustration and examples of unidirectional and bidirectional scale formats.

Form A:				
A. Poor	B. Fair	C. Good	D. Very Good	E. Excellent
Form B:				
A. Poor	B.	C.	D.	E. Excellent
Form C:				
1. Poor	2. Fair	3. Good	4. Very Good	5. Excellent
Form D:				
1. Poor	2.	3.	4.	5. Excellent

Figure 4. The four scale response formats compared by Frisbie and Brandenburg.

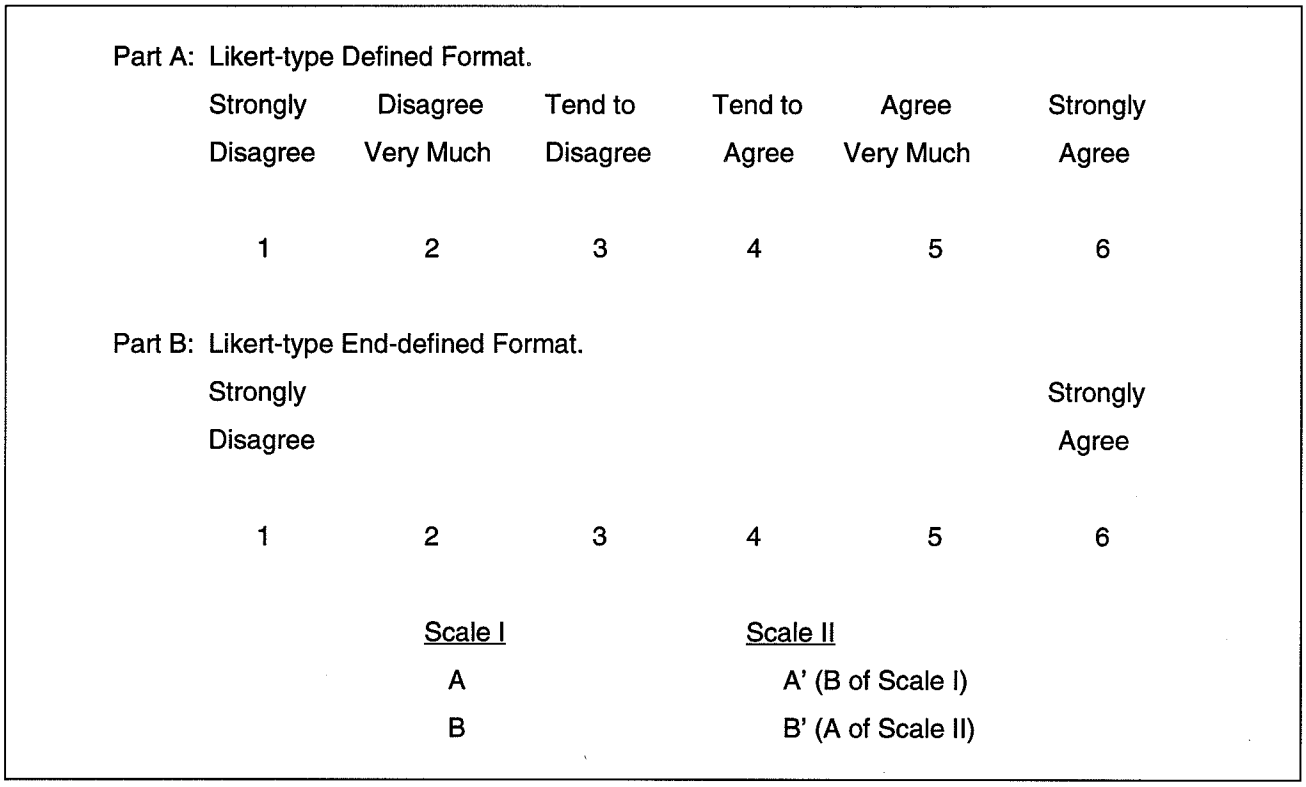

Figure 5. Scale response formats compared by Dixon, Bobo, and Stevick (1984).

part of items in Scale I was switched to the other format. Examples of the two formats are provided in Figure 5.

The results of this study suggested that there was no difference in effect between those two formats. Neither was found to be preferred to the other. The only difference was that significantly more items had a larger standard deviation in the end-defined format than in the defined format. Lam and Klockars (1982) also suggested that there was no significant difference between the two formats, and that the distribution of scores might be directly and predictably influenced by the particular labels used to mark the intermediate response options.

Graphic Scales

Numerous variations in this type of format have been used. According to Guilford (1954), a common feature is that a straight line is displayed with various cues combined with it to aid the rater. The line can be segmented into a variable number of units, or it can be continuous. It can be placed horizontally or vertically. The examples from Guilford (1954, p. 265) are shown in Figure 6.

Landy and Barnes (1979) investigated a different type of graphic rating scale, in which raters were requested to decide how much of the dimension described was present and to assign a scale value from 1 (minimum amount) to 7 (maximum amount) to each statement. This type of graphic scale is sometimes referred to as the "Thurstone-type scale" (Dawis, 1987, p. 483; Latham & Wexley, 1977). Two or three items with the lowest variabilities are selected on the basis of their average scale values to represent each scale point and are arranged in random fashion.

Standard Scales

Like the scales for assessment of handwriting skills, standard scales provide some standard specimens or examples of performance ability that have previously been calibrated on a commonly used scale of quality. With a set of specimens, the rater can equate a new sample of performance to one of the standards, or judge it as being between two standards. The standards can be key persons about whose behavior the rater already knows (the Man-to-Man Scale), a set of verbal portraits (the Portrait Matching Scale) or actual performance samples as illustrated in Figure 7.

Ratings by Cumulated Points

With this scaling format, an object or an individual is assessed by the sum of scores derived from a number of ratings, whether they are weighted or unweighted. This format is commonly used in check-list scaling methods. The items in weighted scales can be rated by the simple weights of -1, 0 and 1, or more varied

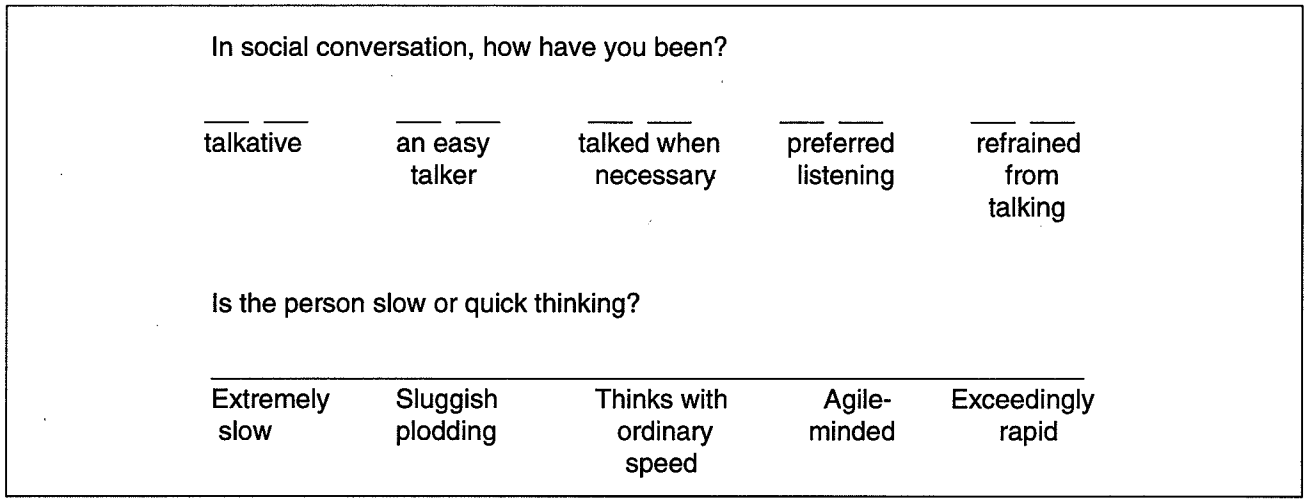

Figure 6. Examples of graphic scales (from Guilford, 1982).

weights (Guilford, 1954). We can use these weighted scales when the emphasis is placed on the intensity of behaviors. For unweighted scales, one can use a checked/unchecked format. In this case, the observations are related to occurrence or non-occurrence, or frequency of behaviors. These two formats are illustrated in Figure 8.

Forced Choice Ratings

For this type of scale, usually two pairs of statements or descriptive adjectives (one pair with high preference value and the other with low preference value) are combined to form an item. The rater is asked to say essentially whether an individual has more of one trait than another of each pair. Figure 9 gives an example adapted from Guilford (1954, p. 275).

Although six different item formats with forced choices have appeared in the research literature, "Form C" (p. 276) is regarded as best. This format includes four statements, all favorable, and the rater selects the two most appropriate statements that describe an individual.

Scale Development

Once items are collected, the most appropriate ones should be selected and arranged according to their relative values. Six techniques are reviewed here, including: Thurstone method, Likert method, Guttman method, Behaviorally Anchored Rating Scales (BARS), Behavioral Observation Scales (BOS), and the Mixed Standard Scales (MSS). The first three were originally devised to assess attitudes, and are called "attitude scale methods" (Guilford, 1954, p. 456). The last three are used to assess job performance skills in various professions other than school settings (hence here they are addressed as "performance scale methods"). Among them, only the BARS method has been applied to the school setting to evaluate university instructors (Jacobs, Kafry, & Zedeck, 1980). But the present authors could not find any research studies with scales employing the BARS or the other two performance scale methods, to assess students' skill performance. However, we assume that they can be effectively modified or applied for the development of scales to rate behavioral deficits or excesses in school settings.

The Thurstone Methods

Thurstone (1928) was reportedly the first to suggest that attitudes can be measured by the opinions individuals endorse as their own (Edwards, 1957; Guilford, 1954). Thurstone (1946) defines an attitude as "the intensity of positive or negative affect for or against a psychological object" (p. 39). In this definition, "positive" and "negative" reflect "favorable" and "unfavorable," respectively, and the term "affect" is synonymous with "feeling" (Edwards, 1957, p. 2).

Thurstone and Chave (1929) described an approach to the construction of an attitude scale. For a scale of attitudes toward the church, they collected opinions from two sources to develop it: (a) Several groups of people and many individuals were asked to write their opinions about the church, and (b) Current literature was reviewed for suitable and representative statements. Thus they obtained 130 statements, which were to be sorted into intervals for the scale.

The criteria for selecting those initial statements were as follows:

1. The statements should be as brief as possible so as not to bore the subjects who would read the whole list for selection.
2. The statements should be either endorsed or

1. A Man-to-Man Scale

The child's letter writing skill is

Excellent	Karen Witte*	18 points
Good	Brett Chave	15 points
Fair	Jamy Thurs	12 points
Poor	John Moore	9 points
Very Poor	Tom Levine	6 points

* All names have been fictionalized.

2. An Anchor of a Portrait Matching Scale.
Oral communication:

J relates simple experiences in chronological order. She properly tells her parents what happened on her way home from school. She coherently retells a short story which she has just read that includes five or six incidents (15 points).

3. Anchors of a Standard Scale Using Performance Samples.

Handwriting Skill:

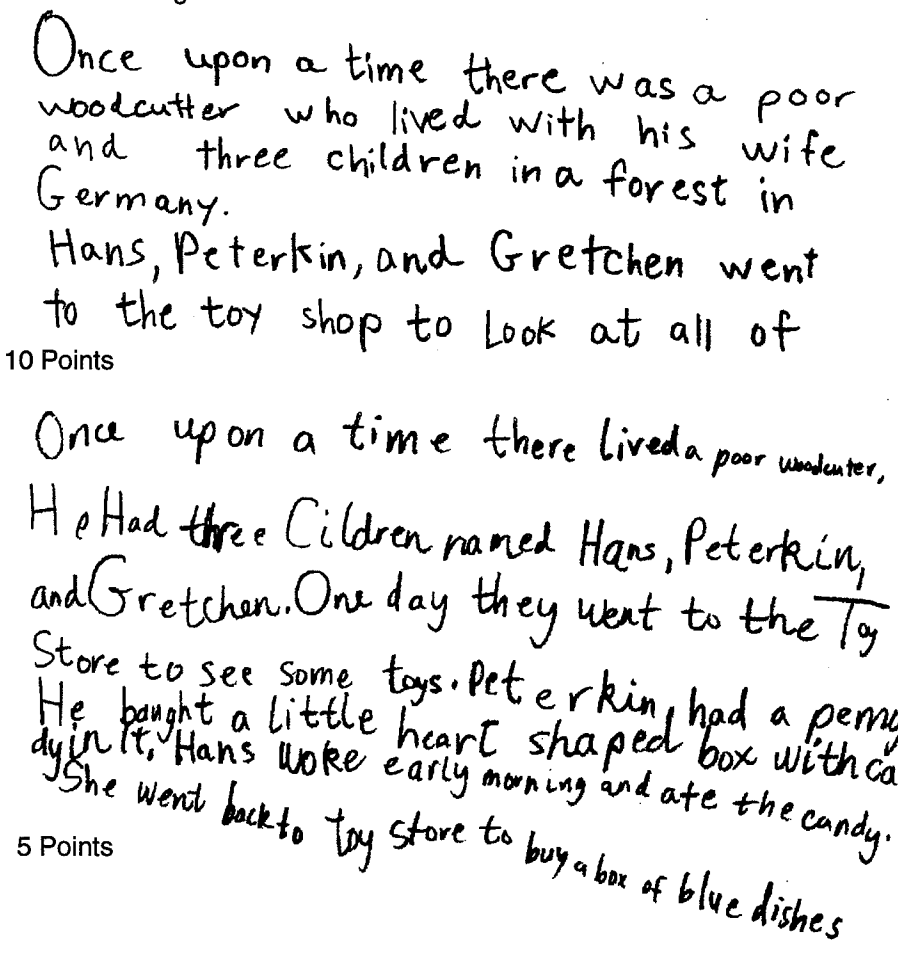

Figure 7. Three types of standard scales.

Checklist for the Descriptive Paragraph

Mark an "X" under each sentence number.

Criteria	Sentences								
	1	2	3	4	5	6	7	8	9
Every sentence about the topic									
Frozen in time									
At least one adjective appealing to a sense									
Complete sentences									
Capitalization									
Punctuations									

Student/teacher opinion on the changes needed: _____

_____

_____

Figure 8. An example of a checklist format.

rejected according to their agreement or disagreement with the attitude of the readers.

3. Every statement should be such that its acceptance or rejection indicates something regarding the reader's attitude about the issue in question.

4. "Double-barreled" (Thurstone & Chave, 1929, p. 37) or ambiguous statements should be avoided unless there seem to be no better neutral statements available. An example of a double-barreled statement was given by Thurstone and Chave (1929, p. 37): "I believe the church has a good influence on the lower and uneducated classes but has no value for the upper, educated classes."

5. At least a fair majority of the statements should really belong on the attitude variable to be measured. In brief, they should be short and to the point.

To anchor response alternatives on the construct continuum, this approach used "the method of equal-appearing intervals" (Edwards, 1957, p. 83). Subjects who had not been involved in the development of the statements were asked to sort the statements, written on separate cards, into a number of intervals. For representing those intervals, cards with the letters *A* to *K* written on them were arranged in order in front of the subjects. Then the subjects were asked to place those statements on the *K* card that seemed to express the most favorable feelings about the psychological construct. Those expressing the most unfavorable feelings were to be placed on the *A* card. For this task, only the middle and the two extreme cards were defined (Thurstone & Chave, 1929). Numerical values are assigned to the eleven intervals with *A* indicating 1 and *K* indicating 11. The configuration of the interval continuum is shown in Figure 10.

The child is

_____ careful
_____ clumsy
_____ tidy
_____ slovenly

Figure 9. An example of a forced choice scale.

The scale value of each statement is determined by the median (or 50th percentile) of the frequency distribution for the statement (Thurstone & Chave, 1929). For example, in Thurstone and Chave (1929), 125 people were asked to assign a number of preliminary statements to appropriate intervals. The resulting distribution for one of the statements was like that in Table 1. The middle number of subjects is $(125 + 1)/2 = 63$. The point at which the 63rd person falls is the median, which lies somewhere between intervals 3 and 4. Therefore, the median can be calculated by the following equation: $Mdn = 3 + (63 - 48)/(73 - 48) = 3.60$. This number should be rounded to the nearest whole number, because all the intervals are represented by whole

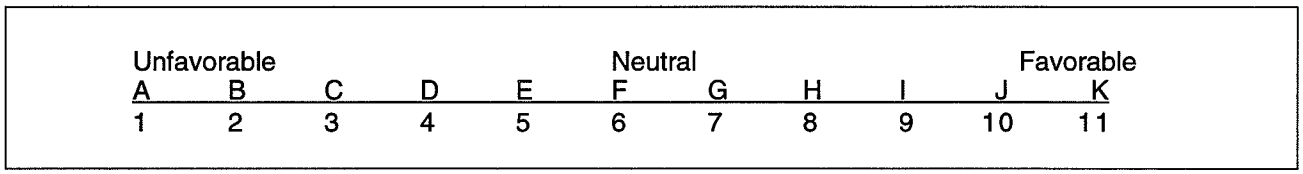

Figure 10. Arrangement of intervals for the Thurstone Equal-interval Continuum.

numbers (Thurstone & Chave, 1929). Hence the term, "equal-appearing intervals" (Edwards, 1957, p. 83).

In addition, the *Q*-value for each statement is computed as the difference between the 75th and the 25th percentiles to indicate the degree of a statement's ambiguity. High *Q*-values are regarded as ambiguous and undesirable for the scale. Among the statements with approximately the same scale values, statements with high *Q*-values are eliminated. Equal numbers of statements are selected from each of the class intervals based on their *Q*-values and the above-mentioned guidelines. About 20 statements are usually selected for an attitude scale so that the scale values of the statements on the psychological continuum are relatively equally-spaced and the *Q*-values are relatively small (Edwards, 1957).

possible combinations. To anchor statements on the scale, the respondent's task is to select one stimulus or statement from each pair on the basis of the scaling dimension, that is, based on which of the two has more of the construct. Thus, all the stimuli or statements are ranked in order and the mean or median of the response frequency of a stimulus is taken as the scale value of that stimulus.

Comparison of the paired stimuli is based on the law of comparative judgment. This law assumes that, for a given stimulus, a most frequently aroused, or modal, discriminal process appears on a psychological continuum. A discriminal process represents the experience or reaction of an individual when confronted with a stimulus and asked to make a judgment of some

Table 1. An Example of Responses to an Imaginary Statement in an Equal Interval Scale Format

Interval	Frequency	Cumulative Frequency
11	0	125
10	0	125
9	3	125
8	7	122
7	7	115
6	15	108
6	20	93
4	25	73
3	23	48
2	17	25
1	8	8
	125	

Respondents to this scale are required simply to respond to it by placing a check mark for each statement that they endorse. Interested readers should refer to Thurstone and Chave (1929) or Edwards (1957).

In another study, Thurstone used the method of paired comparisons with his "law of comparative judgment" (Thurstone, 1927, p. 273). This can be applied to construct a rating scale with a small number of statements (not more than 20). With this method, every two statements in the preliminary pool are paired in all

attribute. The scale difference between the discriminal processes of two specimens or stimuli (discriminal difference) indicates which one of the two specimens better represents the psychological construct. Suppose that the discriminal processes corresponding to stimuli *A* and *B* are denoted as *a* and *b* respectively. If the discriminal difference (i.e., *a* - *b*) is positive, *A* is regarded as better than *B*, and if it is negative, then *B* seems to be better (Thurstone, 1927; Edwards, 1957). This law is usually described by means of an equation

and interested readers should consult Thurstone (1927) or Ghiselli, Campbell and Zedeck (1981).

Summary and Evaluation

Thurstone's equal-appearing interval method is based on the assumption that every statement chosen for the scale is placed in one of the 11 intervals on the psychological continuum, and that the distance between successive intervals is equal. And the scale format for this approach does not require respondents to respond to every item, but simply to check the ones they endorse as their own. This method requires a large number of items to achieve desirable reliability and lacks good indices of validity such as an item analysis (Guilford, 1954).

For the paired comparison method, the items are presented to the rater in all possible combinations and the rater is required to decide which of the pair has more quantity of the defined construct. This process is laborious and therefore does not seem to be widely used.

The Likert Method

Likert (1932) proposed a method of selecting and developing statements for attitude scales based on the following criteria:

1. All statements should be expressions of desired behavior and not statements of fact. Each statement does not measure some past attitude, but the present attitude of the subject. For a statement to involve desired behavior, it is recommended to use the term "should".
2. Each proposition or statement should be stated clearly, concisely, and straightforwardly and should avoid every kind of ambiguity.
3. Generally, it is desirable to word each statement in such a way that the modal reaction to it may be approximately in the middle of the possible responses.
4. It is desirable to have the different statements so worded that about one-half of them have one end of the attitude continuum corresponding to the left or upper part of the reaction alternatives and the other half have the same end of the attitude continuum corresponding to the right or lower part of the reaction alternatives.
5. If multiple choice statements are used, the different alternatives should include only a single attitude variable, not several. (pp. 44-46)

Based on these criteria, more statements should be selected than are likely to be used; in the process of item

1. Should the United States participate in the Vietnam War?

Yes	?	No
(2)*	(3)	(4)

2. Should the United States do something to secure human rights in other countries?

Yes	?	No
(4)	(3)	(2)

3. Should we fight for our country whether it is in the right or in the wrong?

Strongly Approve	Approve	Undecided	Disapprove	Strongly Disapprove
(1)	(2)	(3)	(4)	(5)

4. A person who loves his fellow man should never be willing to engage in a war, no matter what the rationale for the war may be.

Strongly Approve	Approve	Undecided	Disapprove	Strongly Disapprove
(5)	(4)	(3)	(2)	(1)

* The number in parentheses indicates the value of each alternative.

Figure 11. Examples of assigning numerical values to response alternatives in the Likert scales (adapted from Likert, 1932).

analysis, inappropriate items will be deleted. Then, for each statement, numerical values are assigned to a cluster of response choices in an arbitrary order. If, for example, a statement has five alternatives, number 5 could be assigned to the one that is arbitrarily considered "favorable" and the number 1, to the one considered "unfavorable." Figure 11 shows some examples of assigning numerical values to response alternatives. Arbitrary numbers can be assigned to those response alternatives; however, the order of the numerical values in a cluster of alternatives plays an important role, rather than the numbers themselves. For items of the formats 1 and 2 in Figure 11, when values other than 2, 3, and 4 (i.e., 1, 3, and 5) are assigned to the same alternatives respectively, the relative position of an individual's score is not changed. In the same manner, when the values 1, 3, 4, 5, and 7 are assigned to the different positions corresponding respectively to 1, 2, 3, 4, and 5 for the items of the formats 3 and 4 in Figure 11, the same results are obtained (Likert, 1932).

Item analyses ought to be conducted by calculating the correlation coefficient of each statement with the entire battery under two conditions: (a) to check objectively if the numerical values are assigned properly, and (b) to see if the statements are "differentiating" (Likert, 1932, p. 48). A negative correlation indicates that the numerical values are not properly assigned, and that therefore the 1 and 5 ends (in items 3 and 4 in Figure 11, for example) should be reversed. A zero or very low correlation means that the statement is undifferentiating, that is, it fails to measure what the rest of the statements measure. Only those items with high positive correlations and those with high negative correlations, but with the numerical values reversed, should be selected for the final score.

Likert (1932) initially used a laborious calculation (i.e., the method of sigma units) for the scoring system, but later tested a simpler rating system with numerical units, as outlined above. The scores obtained by these two methods "almost perfectly" correlated (Likert, 1932, p. 26), justifying the use of the simpler method in which a total score for each individual is obtained by summating his/her scores across individual items. Hence the term "the method of summated ratings" (Bird, 1940, p. 159).

Summary and Evaluation

The Likert method depends primarily on item analyses for validity of items and item selection. Unlike the Thurstone method, it uses a scale format in which raters respond to every item by deciding how much of a construct it has. If one intends to use a summation score across items, this approach is recommended.

The Guttman Method

This method is also referred to as the "scalogram" analysis or the cumulative scale method (Edwards, 1957). For descriptions of this method, readers should refer to Edwards (1957) and Guilford (1954). Guttman (1944) believed that a genuine scale exists when homogeneity or unidimensionality is virtually complete. For a unidimensional scale, all the statements that constitute the scale must be related to the attributes that define the concept or behavior to be measured, so that they should be calibrated on the same continuum to measure a common content. The scale must guarantee the unidimensionality in such a way that an individual with a higher rank or score than another individual on the same set of statements must also rank just as high or higher on every statement in the set. This condition is what is called perfect reproducibility (Edwards, 1957).

Procedure

In this method, a large number of descriptive statements defining a construct (the "universe") is collected and a reasonably small number of them (20 or less) is selected, based upon of intuition and experience (Edwards, 1957). Each of these selected items has two (or more) response choices, e.g. "agree" and "disagree." A group of individuals responds to every item in such a way that, if they agree on the item and choose the agree category, we assign the weight 1 to the response. If they respond to the item by choosing disagree, we assign the weight 0 to it. On the basis of the total scores across items, the respondents are placed in a matrix (scalogram) in rank order. They are listed in a column, and in each row of the matrix are the item columns in presumed order of relation to total score (see Table 2).

The next step is to set a cutting point on each response column. The cutting point should be located so that (a) Error is minimized, and (b) No category has more error than non-error (Guttman, 1947). Any 1-point responses that fall below the cutting point and any 0-point responses that fall above it are errors. A cutting point for a response column is indicated by a horizontal line just below it. Statement 4 in Table 2, for example, includes three errors in all. If, however, the cutting point was set below subject 10, the result would be four errors (3 above the cutting point and 1 below it) instead of three. Both of the response categories 1 and 0 have more non-error responses than errors, reflecting an appropriate cutting point. So the cutting point for statement 4 has met the two conditions.

The response sample in Table 2 includes a total of 80 responses (20 X 4) and 11 errors. The proportion of errors to the total responses is 11/80 = .138. Therefore the proportion of non-error responses (the coefficient of reproducibility) is 1 - .138 = .863.

If the coefficient of reproducibility is .90 or greater, the set of statements is scalable (Edwards, 1957). The patterns of responses must also indicate substantial frequencies for scale types as opposed to non-scale types (Edwards, 1957), thereby signifying consistency

in response types. For example, the cutting point for Statement 4 in Table 2 falls between Subjects 5 and 6. Take the first 5 subjects (i.e. Subjects 1-5). For Statements 1 and 4, all of them responded by choosing Category 1. For Statement 2, three out of five subjects chose 1 and the remaining two chose 0. For Statement 3, four subjects selected 1. Thus, the overriding response pattern across the statements is 1 1 1 1, which is the scale type. Either 1 0 1 1 (i.e. Subjects 3 and 4) or 1 1 0 1 (i.e. Subject 5) is a non-scale type.

Table 2. The Guttman Method as Applied to an Example Scale

	STATEMENTS								
	1		2		3		4		
SUBJECTS	1	0	1	0	1	0	1	0	SCORE
1	X		X		X		X		4
2	X		X		X		X		4
3	X			X	X		X		3
4	X			X	X		X		3
5	X		X			X	X		3
6	X		X		X			X	3
7	X		X		X			X	3
8	X			X	X			X	2
9	X			X		X	X		2
10		X		X	X		X		2
11	X			X	X			X	2
12	X			X	X			X	2
13	X			X	X			X	2
14	X			X		X		X	1
15	X			X		X		X	1
16		X	X			X		X	1
17		X		X		X	X		1
18		X		X	X			X	1
19	X			X		X		X	1
20		X		X		X		X	0
frquency (f)	15	5	6	14	12	8	8	12	
proportion (p) & q (=1-p)	.75	.25	.3	.7	.6	.4	.4	.6	
errors (e)	1	1	1	2	1	2	3	0	Total = 11

To assign scale values to the statements, we should use the following procedure (Guttman, 1944):

1. For each statement in Table 2, a bar is provided with the agree category on the right side and the disagree category on the left side, as shown in Figure 12. The areas that the categories occupy should be adjusted according to their proportions.

2. These bars, representing the data for the statements, should be arranged in a column in such a way that the statement with the smallest proportion for the agree category is placed on the top and the statement with the largest proportion at the bottom.

3. The dividing line between the categories on each bar is extended (in a dotted line) across the other bars, and the proportions of the areas between those extended lines should be calculated and written below the bar at the bottom.

4. Finally, for scale values, ordinal numbers beginning with 0 are assigned to those areas from left to right. Thus, Figure 12 shows that the resulting numerical values for the statements 1, 2, 3, and 4 are 1, 4, 2, and 3, respectively.

Summary and Evaluation

The Guttman method is a scale analysis procedure rather than a scale construction technique (Edwards, 1957; Guilford, 1954). It presents an excellent system to monitor the logical nature of responses and consistency of rating, but has some weaknesses. Guilford (1954) summarized these criticisms as follows:

> 1. Even when total scores reach an acceptable level of reproducibility, the scalability criterion is

hard to achieve.

2. Even when the criterion is achieved, we cannot be sure whether the scale has unidimensionality or it has more than one variable.

3. Response popularity produces reproducibility in such a way that responses piled up in one category result in high producibility.

Behaviorally Anchored Rating Scales

Smith and Kendall (1963) incorporated the Critical Incident Technique proposed by Flanagan (1954) into their unique retranslation procedure and marked the beginning of an influential method of behavior rating scale construction: Behaviorally Anchored Rating Scales (BARS). This method was designed to provide a scale in which head nurses could evaluate the job performance of their subordinate nurses. Although this purpose, namely the evaluation of job performance, is similar to the Critical Incident Technique, which purported to evaluate aviation skills, the examples used by this method did not represent actually observed critical behaviors but "inferences and predictions from observations" (p. 150). Because the BARS are constructed based on expectations of a behavior, the term "behavioral expectation scales (BES)" is used synonymously with BARS (Bernadin, 1977; Latham, Fay, & Saari, 1979).

The original BARS procedure proposed by Smith and Kendall (1963) is highly systematic and has well-designed psychometric properties. It has a major weakness as well: it is laborious and time-consuming. Some variations of this procedure have also been suggested to reduce the time factor. Other weaknesses also exist such as scarcity of items for midrange performance levels and inconsistent reports from studies on its interrater reliability. Three procedures for BARS are presented below: the traditional BARS, the shortcut BARS, and the optimal procedure.

STATEMENTS	DISAGREE				AGREE
2		70%			30%
4		60%			40%
3	40%				60%
1	25%				
frequency	25%	15%	20%	10%	30%
score	0	1	2	3	4

Figure 12. Assignment of scale values to the statements in a scale by the Guttman method.

The Traditional BARS Procedure

The traditional BARS procedure includes three steps: (a) the critical incident procedure, (b) retranslation, and (c) scaling (Smith & Kendall, 1963). The whole procedure requires two groups of people with sufficient knowledge and experience on the job to be investigated: one group for collection of statements or items, and the other group for retranslation and scaling.

The Critical Incident Procedure

Statements are first collected using the critical incident technique. Of the two groups of participants, one group lists performance qualities of the target job and describes specific illustrations or examples representing definitions of high, low, and acceptable performance for each quality. The same group clusters these examples into smaller sets of performance dimensions or qualities which they typically define.

Retranslation

The second group of participants (judges) are instructed to reallocate ("retranslate") the critical examples to the same set of dimensions as in the previous step. This retranslation step aims at eliminating inconsistent and inappropriate examples or qualities. Typically, an example is retained if a certain percentage (usually 50-80%) of the group assigns it to the same dimension as did the first group (Schwab, Heneman III, and & DeCotiis, 1975). Qualities are also eliminated if examples are not consistently reassigned to those qualities to which they were assigned by the first group.

Scaling Examples

Smith and Kendall (1963) proposed formatting a series of continuous graphic rating scales, whose anchors are arranged vertically. A set of 7- or 9-point graphic scales is typically used for the BARS procedure (Schwab et al., 1975). Latham, Fay, & Saari (1979), however, introduced a summated scale for this method.

They provided a 5-point Likert-type scale for each item, using fixed standard scaling, which predefined a certain percentage of occurrences for each scale point (see Figure 14). Such a scale, however, could result in serious errors in the levels of satisfactoriness (Bernadin & Smith, 1981).

Using such a graphic scale format, the second group is asked to rate the behavior described in each example in terms of how effectively and ineffectively it represents performance on a specific dimension. The average (mean) rating assigned to an example determines the degree of performance effectiveness of the example. At this stage, those examples with standard deviations of more than 1.50 are eliminated. The final instrument is a series of vertically arranged graphic scales, one for each dimension. Those scales are anchored by the examples that meet the retranslation and the standard deviation criteria. Usually one scale consists of 6 or 7 examples (Schwab et al., 1975).

Summary and Evaluation

For developing rating scales, the retranslation procedure has received considerable attention as a method of identifying performance dimensions and corresponding behavioral examples (Dickinson & Zellinger, 1980). The greatest advantage of this procedure, along with the critical incident technique, is that future users of scales play a major role in developing and examining them (Smith & Kendall, 1963). In its essence, this approach was designed to encourage and standardize direct observation of behaviors, which may familiarize raters with summary ratings for future scale use (Bernadin & Smith, 1981). Also, it can enhance the reliability and validity of the scales, because the contents and the actual words of the items derive from the rater's participation in the process of scale construction. Green, Sauser, Jr., Fagg and Champion (1981) reviewed and summarized studies on BARS as follows:

1. BARS have medium to high reliability and seem to have adequate convergent validity, correlating highly with appropriate objective measures.
2. Field tests have demonstrated that BARS are superior to typical graphic rating scales in terms of reliability, validity, and freedom from halo and leniency errors. (For the concepts of "halo effects" and "leniency errors," see the section on the Evaluation of Scales.)

Despite all these advantages, the procedure has some problems as well. BARS constructed by this procedure generally lack the anchors for the midrange of performance effectiveness, because items assessing the midrange are most likely to be rejected. Because of the rejection of a large percentage of items, this procedure is extremely wasteful and costly in terms of required hours (Green, Sauser, Jr., Fagg, & Champion, 1981). These major problems delimit the applicability of the BARS procedure.

Zedeck, Imparato, Krausz, and Oleno (1974) examined the effects of participants' organizational levels on the behavioral incidents and the behavioral dimensions defined by the incidents in the behaviorally anchored rating scales. Two groups of nurses participated in constructing BARS which would be used to assess the nurses' job performance. One group was composed of head nurses representing the "supervisory level" and the other group of registered nurses representing the "subordinate level." The two groups independently developed BARS by the traditional BARS procedure. The two organizational levels identified similar behavioral dimensions, but the behavioral incidents defining dimensions were valued differently. Presumably, this difference is a result of differences between the subordinate group and the supervisory group, in terms of the behavior believed to be necessary for adequate performance. If further studies replicate the same comparison and corroborate the results, this suggests that a BARS to be used by one organizational level should be constructed by the same level. It is usually assumed that critical incidents used to anchor the scales should represent the same level of effectiveness for all raters (Smith & Kendall, 1963).

Schwab et al. (1975) reviewed several studies on BARS and concluded as follows:

1. Leniency effects were not high, although they were inconsistent across results of studies.
2. The results of these studies do not support the high promise regarding independence of scales representing dimensions.
3. The results suggest that BARS are not a panacea for high interrater reliability.

Jacobs, Kafrey, and Zedeck (1980) also made a comprehensive review of BARS literature and reported that BARS research had focused mostly on the quantitative rather than on the qualitative criteria. They suggested that BARS may have greater potential when assessed on the utilitarian and qualitative criteria, although it is no better or worse than other methods when assessed on a quantitative basis. For concepts of these criteria, see the Scale Evaluation section.

The Shortcut BARS

Because the traditional BARS procedure is extremely time-consuming, Green, Sauser, Jr., Fagg, and Champion (1981) attempted another method for developing BARS cheaply and efficiently, while retaining rater participation in the process. The final product of this method was to be used by university students to evaluate their instructors. It requires only one group of raters and does not include the retranslation procedure; rather, sufficient rater training is accomplished

through models and examples. For this approach, Green et al. used two sessions: (a) the individual session, and (b) the group session.

The Individual Session

A model BARS is presented to a group of participants who have had sufficient access to the performance of a job, to illustrate necessary ingredients of a well-constructed scale. Then they are given eleven criteria:

1. The scale should be anchored by behavioral descriptors rather than adjectives and numbers.
2. The descriptors should be real incidents of behavior that might actually occur in the work situation.
3. The descriptors should be clear and to the point.
4. The descriptors should be phrased in enough detail that the rater can easily understand them.
5. The descriptors should be phrased in "expectation" format (p. 766) (i.e. "should" rather than the past form of a verb).
6. The descriptors should be placed properly on the line so that each step higher represents a real increase in performance.
7. The spacing should represent the real perceived distance between descriptors.
8. There should not be any large gaps along the scale.
9. The descriptors should describe a sufficient variety of different situations.
10. There should not be any descriptors which do not really belong in this category (dimension).
11. A rater should be able to use the scale without difficulty. Based on these criteria, the raters criticize an example scale and then they are offered feedback. A poorly constructed BARS is also given for them to criticize in terms of the same criteria. (Green et al., 1980, p. 766)

A number of pre-established dimensions (e.g. preparation, delivery of instructional contents, responding to students' questions, etc.) are provided, and each rater is assigned to one of them. The raters individually write 20 behavioral anchors concerning the assigned dimensions. Then each rater is asked to place these behavioral descriptors along the assigned dimension, e.g. from most to least, excluding poor ones and writing additional ones if necessary to end up with 9 to 11 properly placed anchors. Next, the raters revise their scales as many times as necessary to be in a satisfactory form. This revision is based on the 11 criteria described above. After this revision, each rater is requested to make a final draft of the scale.

The Group Session

The same individuals who participate in the individual session can also work in a group. They write criticisms of each of the individually developed scales in terms of the same eleven criteria as in the previous session. Also they are instructed to develop one BARS that includes the best points of the scales. They are allowed to add new behavioral anchors to improve their BARS. Finally, they are asked to review the scale they have constructed in terms of the eleven criteria and to make additional appropriate revisions.

Summary and Evaluation

Green et al. (1981) designed this procedure to construct a BARS without the major weakness (time and efforts) of the traditional BARS procedure. For the shortcut procedure, they substituted the retranslation with rater training, and used only one group of raters. They also compared the shortcut-derived BARS with those derived by the traditional BARS procedure. The convergent validity (the correlation between the two measures) was very high for all the scale dimensions. The shortcut BARS was more susceptible to leniency errors, while the traditional BARS was more susceptible to halo effects. However, experts could not distinguish among the BARS constructed by the two methods. This study also reported that the reliability of the BARS constructed by groups of students was slightly higher than that of either the traditional or the individually-derived BARS, while the convergent validity of the individually-derived BARS was higher than that of the BARS developed by the group.

Thus Green et al. (1981) and Champion, Green and Sauser, Jr. (1988) seem to have regarded the two scales produced by the two sessions as separate BARS. Obviously the BARS constructed by the group session was derived from those by the individual session. Therefore, the former should not be treated as completely independent of the latter, but rather should be regarded as the replacement for a part of the retranslation procedure in the traditional BARS method. Since the individually-derived BARS better represents the shortcut BARS, the shortcut BARS procedure is characterized by using only one group of raters and the retranslation procedure is absent (Champion, et al., 1988).

One of the advantages of the shortcut BARS procedure is that it produces scales of almost equal quality to those developed by the traditional BARS procedure, while retaining rater participation in the process. Another advantage is that it costs far less in terms of time and effort than the traditional method. Champion et al. (1988) supported these results, suggesting that the shortcut BARS might be a substitute for traditional BARS with little loss in measurement quality.

The Optimal Procedure

Bernadin, LaShells, Smith and Alvares (1976) examined the effects of different developmental procedures and formats for BARS on various psychometric criteria such as reliability, validity, and halo effects. Based on their findings, they made the following rec-

ommendations:

1. Have one group of participants place the critical incidents into appropriate dimensions. After eliminating items that do not meet a 60% placement criterion, organize the appropriate items into their placed dimensions and have another group rate them on their desirability to obtain more stable items. For example, assume that a group of teachers presented a number of statements as critical incidents for children's adjustment behavior. They should classify those statements into some dimensions such as Self-Adjustment, Home Adjustment, Social Adjustment, School Adjustment, and Physical Adjustment, as in the case of the Child Behavior Rating Scale (Cassel, 1962). If less than 60% of the teachers placed a specific item or statement into a specific dimension, the item should be eliminated. The researcher organizes the remaining items into the same dimensions as the teachers did. Then another group of teachers (judges) examines the items on their desirability.

2. To inhibit leniency error (i.e. the tendency for raters to assess individuals too high or too low) and increase discriminability across raters (i.e., the degree to which we can distinguish between consistent and inconsistent raters), the raters should develop dimension clarification statements to be placed at anchor points on each scale.

3. To decrease leniency error, increase rating variability across dimensions and within ratees, and increase discriminability across raters, the raters should place observed critical incidents directly relevant to the ratee at the positions on the scale where they seemingly belong. They should then compile a summary rating by averaging the scale values of the new items.

Kinicki and Bannister (1988) incorporated these recommendations into their scale construction procedure to examine two fundamental assumptions underlying BARS:

1. The BARS provides a vehicle for helping raters to identify specific behaviors that describe effective and non-effective performance.

2. The critical incidents used to anchor the scales represent the same level of effectiveness for all raters.

Procedure

The first group of participants (undergraduate students) generated and defined the dimensions of teaching effectiveness. Then they were requested to illustrate behavioral examples on the basis of what was judged to be good and poor performance on each dimension.

The second group was given the list of performance dimensions along with their definitions and a randomized list of behavioral examples. The participants of this group were asked to independently sort the examples into the dimensions best represented by those examples. This process is equivalent to the retranslation. Items were eliminated if they did not meet the 60% retranslation criterion.

With the surviving items, the third group of raters independently rated the level of effectiveness of each behavioral example on its dimension on a 7-point scale. Items with standard deviations greater than 1.50 were eliminated.

To obtain a measure of the behavioral specificity of each item, the fourth group independently rated each item along two scales, illustrated in Figure 13. Ratings from these two scales were summed to represent such a measure.

Summary and Evaluation

This procedure is basically intended to retain the traditional BARS procedure, while enhancing psychometric properties of the scale. Kinicki and Bannister (1988) reported a considerable variation in the specificity of the behavioral anchors, which resulted in high interrater variability. The results also suggested that the evaluation of the critical incidents was significantly different across rating contexts. To reduce the interrater variability, the authors have proposed that the raters be trained to observe and document specific examples of performance. Jacobs et al. (1980) performed a comprehensive review of BARS literature and summarized findings up to that time : (a) The BARS procedure is no better or worse than other methods when evaluated on a quantitative basis, but (b) It has greater potential when evaluated on the utilitarian and qualitative criteria.

Behavioral Observation Scales

Latham and Wexley (1977) first developed the procedure of Behavioral Observation Scales (BOS) to assess the job success of logging supervisors. The procedure of BOS is similar to that of BARS in three ways (Latham & Wexley, 1977): (a) Both methods incorporate the critical incident technique as a means of collecting items, (b) Items for both BARS and BOS are worded in the terminology of future users as a result of their participation in the process of scale construction, and (c) Both procedures take into account the multidimensionality or complexity of performance.

However, they are different from each other in some ways.

1. The development of BARS is similar to the Thurstone method. Items with numerical values that represent the effectiveness of performance, are arranged on a continuous graphic rating scale. The BOS is closer to the Likert method, in that individuals are observed and rated on a 5-point scale, based on the frequency with which they engage in the behavior described by each statement (item).

2. As a means of validating items, the BARS procedure uses the retranslation (except for the shortcut

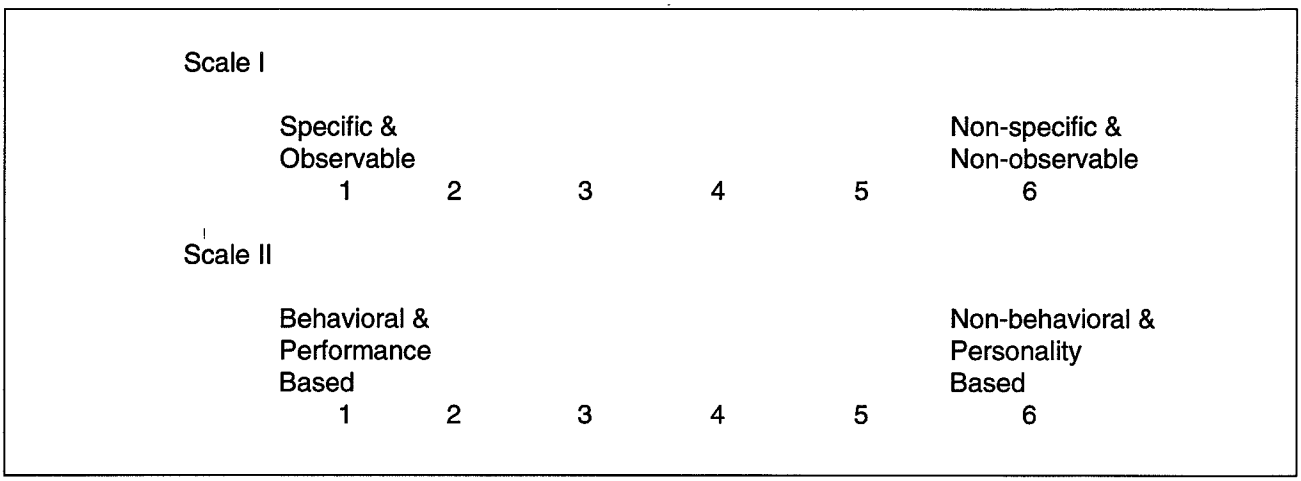

Figure 13. The two scales used by Kinicki et al. (1988) to obtain a measure of the specificity of items.

BARS), while the BOS method conducts item analysis, as in the Likert method, by calculating the correlation coefficient of each statement with the entire battery.

Procedure

For precise observation and assessment of frequency, the target behavior should be clearly defined. Then items are collected by the critical incident technique. In interviewing participants for the items, three points are emphasized.

1. What are the circumstances surrounding a specific incident?

2. On the basis of observation, describe what the individual exactly did that was effective and ineffective so that vague traits or attributes are documented in terms of overt action.

3. How is an incident an example of effective or ineffective behavior? Among the incidents thus collected, those which describe essentially the same behavior are grouped into one cluster, which is to be a behavioral item. Then similar clusters are grouped together to form a BOS. The scales may take the form of a dichotomous (checked/unchecked) rating, but a Likert-type rating format is recommended for more accurate assessment of frequency.

With these scales, the raters independently observe an individual for a specified period of time and rate each individual's performance on a 5-point Likert-type scale. Proportions of frequency and their matching numerical values on the scale are illustrated in Figure 14.

To select most discriminating items, the item analysis used for the Likert method is employed. Items with coefficients of +.30 or smaller should be eliminated. A total score for each individual is calculated by summing the rater's responses to all the items.

Summary and Evaluation

The BOS procedure is similar to the BARS method in that it includes participation of future users in scale construction. The BARS, however, is based on the rater's expectations of performance of a specific task, while the BOS is derived from direct observations of the performance.

The BOS procedure has been less investigated than the BARS procedure. Latham et al. (1979), however, suggested some advantages of using BOS for performance appraisals, some of which include:

1. Misunderstanding of the items in the BOS is minimized because, like BARS, future users of the scales play a major role in scale construction, supplying items in their own terminology.

2. The BOS is content valid, because either the construction procedure or the actual rating on the established scales is based on close observation and thorough evaluation of an individual's performance.

3. The BOS can be used alone or as a supplement to performance descriptions because they consist of specific statements of what behaviors are required of an individual in a given task.

4. The BOS provides performance feedback to motivate individuals to make positive behavior changes.

The BOS has a couple of major disadvantages as well: (a) time constraints, and (b) sample size. Qualitatively, a series of BOS can be developed immediately after the critical incidents are collected, but quantitatively, it requires a sample size of several hundred people and considerable item analyses (Latham & Wexley, 1977).

The choice between BARS and BOS procedures is reduced to a preference for Likert or Thurstone scales (Latham, et al., 1979). The BOS may be preferred when a high degree of contact happens between the rater and the individuals whose behaviors are to be appraised. On the other hand, the BARS may be preferable when such contact is minimal (Latham & Wexley, 1977).

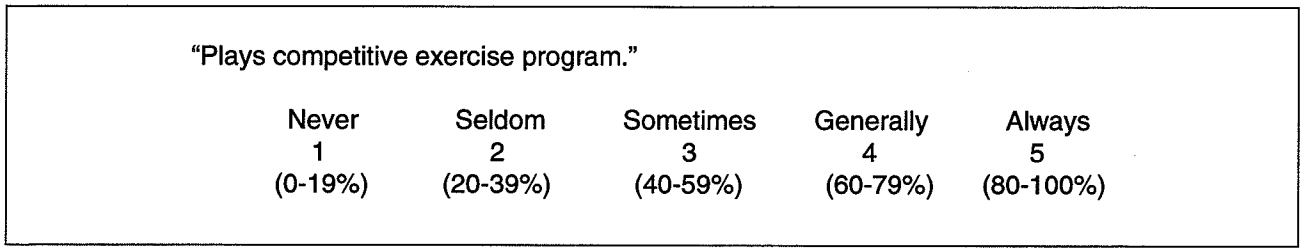

Figure 14. An example of a Likert-type scale for the BOS procedure.

Mixed Standard Scales

Blanz and Ghiselli (1972) proposed a "new merit rating method" (p. 185), namely the mixed standard scale (MSS), which was developed in Finland and applied to various occupations. It has minimized the common errors in rating, namely halo and leniency errors and provided a useful index of the accuracy of rating (Blanz & Ghiselli, 1972). By and large, the characteristics of this method are very close to those of the Guttman method.

Procedure

Blanz and Ghiselli (1972) did not specify any means of collecting and choosing items. But Dickinson and Zellinger (1980) reported that the greatest amount of discriminant validity ever reported had been achieved by using the retranslation procedure developed by Smith and Kendall (1963) in conjunction with the MSS format. For the term "discriminant validity," please refer to the Quantitative Criteria section. Inferentially, the MSS method can be combined with the critical incident technique to enhance content validity. Actually, Saal and Landy (1977) generated their MSS from those items, developed consistently with the guidelines proposed by Smith and Kendall (1963).

After collecting items, the rater is presented with the MSS format, which consists of three behavioral statements for each of the several dimensions or "traits" (Blanz & Ghiselli, 1972, p. 188) of performance to be rated. For each dimension, one of these three statements represents an example of superior performance, another describes an example of average performance, and the last an example of inferior performance (Blanz & Ghiselli, 1972).

All the statements are mixed in a random order to disguise the identity of the performance dimensions and the order-of-merit continua underlying them. After completion of the rating form, the statements and the responses to them are rearranged in order of superiority (i.e. Statement I being the best description, and Statement III the poorest in Tables 3 and 4). The rater is required to respond to every statement instead of selecting the one best describing the person to be rated. The response to each statement must be coded in one of the three ways: (a) "0" for "fits the ratee," (b) "-" for "the ratee is poorer than the statement," and (c) "+" for "the ratee is better than the statement." For every sequence of responses to the three statements within a scale or dimension, a numerical value is assigned according to the degree of the trait. The response types with numerical values include both logical and illogical responses as in Table 4. For example, the response "- 0 +" is logical and "faultless" (Blanz & Ghiselli, 1972, p. 187). On the other hand, the response "+ 0 +" is illogical, because a person who is better than a superior behavior cannot fit an inferior behavior. The logical and illogical types of responses are closely related to the scale and non-scale types in the Guttman scale. All the logical sequences and their points are illustrated in Table 3.

Blanz and Ghiselli (1972) also assigned numerical points to illogical sequences of responses (p. 189). Saal and Landy (1977), however, indicated that two of the possible illogical rating combinations, namely "0 0 +" and "- 0 0", had been omitted from consideration by Blanz and Ghiselli (1972). Saal (1979) suggested an alternative coding system including the omitted combinations. The comparison of the original and the revised systems is shown in Table 4.

The original system of allocating numerical points to the illogical combinations is illogical or inconsistent, because no rules were suggested for them. On the contrary, for the revised system, clear-cut criteria are provided. Points for the three types of responses to each of the three behavior levels, namely superior behavior (SB), average behavior (AB) and inferior behavior (IB), are indicated in Table 5. Based on these points for the responses to the individual statements, a point for a combination of the three responses within a dimension is calculated by the following equation: scale point (p) = SB + AB + IB - 8.

Saal (1979) found that no difference exists between the original coding system and the proposed revised system in terms of leniency and halo errors, and interrater reliability. However, the revised system is

recommended because: (a) A more consistent coding system may facilitate greater interrater agreement, and (b) Improving the "face validity" of the coding system may mitigate a common objection to the MSS format, namely lack of rater (and ratee) acceptance (p. 427). The term face validity concerns the extent to which an instrument "looks like" it measures what it is supposed to measure. If potential users like the types of items in an instrument, the instrument can be said to have face validity (Nunnally, 1978).

Summary and Evaluation

The most remarkable feature of the MSS method is its scoring format. The logic of the scoring system is similar to that of the Guttman method: any illogical responses of the rater are regarded as errors, and the error counts help to identify erroneous raters and ambiguous dimensions, a useful indicator of reliability (Blanz & Ghiselli, 1972). If the errors are unique to a particular rater, they can be attributed to the rater's lack of training. Yet if a certain error is consistent across raters, the dimensions for the scales should be examined for any ambiguity. Errors can also occur when an individual's performance is not appropriately known to the raters, or when a great deal of inconsistency exists in an individual's behavior (Blanz & Ghiselli, 1972).

The MSS is also characterized by its disguised items and dimensions, which reduce halo and leniency errors. Disguising the continuity of the scale on which behavioral expectations are presented, however, does not seem to improve rating in terms of psychometric properties (Arvey & Hoyle, 1974; Finley, Osburn, Dubin, & Jeanneret, 1977).

The MSS, compared to the BARS, produced as much discriminant validity and less method bias (Dickinson & Zellinger, 1980). Research also supported the expectations made by Blanz and Ghiselli (1972) that the MSS rating would reduce the halo effect and the leniency errors (Saal & Landy, 1977).

Scale Evaluation

We can construct rating scales by one or a combination of the procedures discussed so far. Once they are constructed, they need to be evaluated so they can be validated psychometrically. Some of validation criteria can vary according to the nature of ratings. For example, with attitude scales, individuals usually rate their own affects or feelings associated with psychological constructs. On the other hand, with performance scales, raters are required to observe the behaviors of individuals whose performance on pre-defined tasks is to be rated, and to make inferences regarding those behaviors. These two different types of ratings can be used for different purposes, and so their utilization criteria can be different.

The nature of performance ratings by informants, not by the ratees themselves, raises two issues: (a) equivalence of situations, and (b) interpretation (Jacobs et al., 1980). Equivalence of situations demands that the situations under which different raters evaluate an individual's behaviors should be comparable. Therefore, evaluation of individuals by one rater under one situation should not be compared to the evaluation of the individuals by a different rater under a different situation. Raters are also required to make inferences

Table 3. Sequences of Logical Responses to the Statements Within a Dimension for the MSS Format

Combination of Statements			Points
I Superior behavior	II Average behavior	III Inferior behavior	
+	+	+	7
0	+	+	6
-	+	+	5
-	0	+	4
-	-	+	3
-	-	0	2
-	-	-	1

Note. From "The mixed standard scale: A new rating system" by F. Blanz and E. E. Ghiselli, 1972, *Personnel Psychology, 25*, p. 188.

Table 4. The Original and Revised Numerical Systems for the Possible Combinations of Responses to the Mixed Standard Scales

	Response Combination			Numerical Ratings	
No.	Superior	Average	Inferior	Original	Revised
1	*+	+	+	7	7
2	+	+	0	7	6
3	+	+	-	7	5
4	+	0	+	4	6
5	+	0	0	3	5
6	+	0	-	4	4
7	+	-	+	3	5
8	+	-	0	2	4
9	+	-	-	1	3
10	0	+	+	6	6
11	0	+	0	6	5
12	0	+	-	6	4
13	0	0	+	omit	5
14	0	0	0	4	4
15	0	0	-	4	3
16	0	-	+	5	4
17	0	-	0	2	3
18	0	-	-	1	2
19	-	+	+	5	5
20	-	+	0	5	4
21	-	+	-	5	3
22	-	0	+	4	4
23	-	0	0	omit	3
24	-	0	-	3	2
25	-	-	+	3	3
26	-	-	0	2	2
27	-	-	-	1	1

Note. From "Mixed standard rating scale: A consistent system for numerically coding inconsistent response combinnations" by F. E. Saal, 1979, *Journal of Applied Psychology, 64*, p. 424. **+: ratee is better than this behavior; 0: ratee is the same as this behavior; -: ratee is worse than this behavior.*

Table 5. Numerical Points for the Three Types of Responses to Each Behavior Level Based on the Criteria Suggested by Saal (1979)

	Responses		
Behavior Levels	+	0	-
Superior behavior (SB)	8	7	6
Average behavior (AB)	5	4	3
Inferior behavior (IB)	2	1	0

regarding what is expected on the rating scales and their relationship to the actually observed behaviors. In this case, the actual element is the "uniformed interpretation of standards, expectations, and forms among raters" (Jacobs et al., 1980, p. 596). This second issue demands sufficient rater training for uniformity. Although these two issues were addressed concerning the BARS, they are also applicable to rating scales developed by other performance (and attitude) scale procedures discussed in the previous section.

Jacobs et al. (1980) proposed three broad categories of criteria as the requisite properties for a scale evaluation system: (a) quantitative criteria, (b) qualitative criteria, and (c) utilization criteria. These three catego-

ries and their subcategories are described below.

Quantitative Criteria

Quantitative criteria are those psychometric properties that are the statistical results of data analyses of behavior ratings. In general, these criteria include: (a) reliability, (b) validity, and (c) accuracy.

Reliability

Reliability refers to the consistency of ratings across conditions. According to the types of conditions, four types of reliability are usually specified: (a) inter-rater reliability, (b) reliability across formats, (c) reliability over time, and (d) internal consistency.

Interrater Reliability. This term refers to the consistency of ratings across raters, or the degree to which ratings on an individual from two different raters tend to converge. This type of reliability is addressed by correlating the ratings by one rater with those by another rater. With self-rating scales such as attitude scales, interrater reliability is equivalent to reliability over time.

Reliability Across Formats. Reliability across formats or indices indexes the degree to which the assessments of an individual from two types of measures are in agreement. It is determined by correlating the ratings on one format with ratings on another format. The most common issue about this type of reliability is the difficulty in construction of two comparable forms of rating scales. Generally, two forms of rating scales are made comparable by selecting parallel or equivalent statements or items for the scales.

Reliability Over Time. Reliability over time is the degree of stability in evaluating an individual at two different points in time. This type of reliability is estimated by the "test-retest" method, that is, by rating an individual's behaviors on the same rating scales at two different points in time, and correlating the two sets of scores.

Internal Consistency. Internal consistency is related to whether or not the statements or items are measuring the same behavior dimension. This type of reliability is addressed by item analysis in which scores on each item are correlated with the scores on the whole battery.

Validity

Generally the term, *validity,* means the extent to which a rating instrument measures what it is supposed to measure. Nunnally (1978) suggested a more operational definition: Validity refers to the extent to which the interpretation of ratings is appropriate, meaningful, and useful in measuring one or more behaviors. We will discuss three broad types of validity: (a) content validity, (b) criterion-related validity, and (c) construct validity.

Content Validity. Content validity means the extent to which the behaviors specified in a set of scale items are related to, and represent, the domain of behaviors they are designed to measure. Therefore, prior to the construction of a rating scale, the target behavior domain should be clearly defined and dimensions (i.e. frequency, intensity, locus, etc.) should be clearly specified. Items constructed by the critical incident technique proposed by Flanagan (1954) will enhance the content validity. Ghiselli et al. (1981) recommended a procedure to enhance content validity, in which two or more panels of experts go through a specific item-examination procedure independently and compare the results. The four-step procedure suggested by Crocker and Algina (1986) to examine a behavior measurement instrument for content validity also includes a panel of experts:

1. The domain of behaviors is clearly defined.
2. A panel of experts is chosen who are capable of making judgments about which items fit into which behavior categories.
3. These experts investigate the degree of match between items and behavior categories using a systematic process. They also make judgments about the extent to which the whole set of items represents all the aspects specified in the definition of the behavior domain.
4. The researcher collects and summarizes the results of this process.

Criterion-Related Validity. Two measures are needed to develop criterion-related validity: (a) a measure of interest under investigation (i.e. the rating scale), and (b) a measure on some standard (the criterion), which may be pre-established. For example, a teacher may assess his or her students on a behavior rating scale at the beginning of a school year to estimate the problems in their classroom behavior which will occur throughout the school year. Or the teacher may rate them on a social competence scale and correlate the scores on this scale with the number of friends whom they socialize with outside school. In these particular instances, the standards are the future problems in the students' classroom behavior and "the number of friends," respectively.

According to whether the criterion and the measure under investigation exist separated in time or at the same point of time, this type of validity usually has two variants: (a) predictive validity, and (b) concurrent validity.

Predictive Validity. This type of validity indicates how accurately a specific test or rating on one behavioral dimension estimates the future occurrence of another behavioral dimension. For example, if a considerable number of students who had scored high on a behavior problem scale later showed high frequency of behavioral problems in the classroom situation, the

scale has a high predictive validity in terms of the criterion, namely behavioral problems in the classroom. Therefore, we can establish the predictive validity of a measurement by: (a) rating an appropriate sample of people on the instrument, (b) obtaining the criterion scores for the same sample of people after a certain period of time, and (c) calculating the correlation between the scores on the instrument and the criterion scores.

Concurrent Validity. Concurrent validity refers to the extent to which we can estimate the performance on one measure (the criterion) using the performance on another measure (the scale under investigation). We usually use a measurement on a rating scale when the measurement on the criterion is highly complicated or time-consuming. We can establish the concurrent validity of a rating instrument by comparing the scores on the rating with direct observations of performance on the criterion. For example, Wilson and Bullock (1989) reported that the Behavior Dimensions Rating Scale (BDRS) had successfully identified approximately 75% of the 1,762 students labeled either behaviorally disordered or normal in a national study. In this case, the scale of interest is the BDRS and the criterion is the classification of the students as either behaviorally disordered or normal.

Construct Validity. Cronbach and Meehl (1955) first coined the term *construct validity* to describe a validation process which had been practiced without any label (Ghiselli, et al., 1981). Unlike criterion-related validity, construct validity is related to the abstract trait of which the measure does not exist in the real world. Those traits are called *constructs*; some examples are sociability, intelligence, self-concept, and academic aptitude. Thus, construct validity refers to the extent to which a rating instrument measures an individual's standing on a construct (Ghiselli, et al., 1981).

Wilson and Bullock (1989) suggested five types of evidence that may be used to examine the construct validity of an instrument, based on the suggestions by Thorndike (1982) and Crocker and Algina (1986):

1. Comparisons are based on the judgment of how the nature of the behavior specified in the items is related to a conception of the construct.
2. A correlation is established between a rating instrument and some other measure of life event that reflects, or is related to, the behavior being rated.
3. Group differences are compared using the scale scores that are expected to vary according to the theoretical bases of the rating instrument.
4. Treatment effect or experimental intervention data are developed which are expected to reflect the behavior in question.
5. Factor analytic data support the interpretation of the scale results through an analysis of the covariance structure of the data.

We recommend that a rating instrument be validated on all or most of the above types of evidence. An abstract construct can be defined by multiple variables and, therefore, it can be assayed by multiple measures on those variables and correlations between the measures (Tindal & Marston, 1990).

In literature we frequently encounter two kinds of validity that are closely related to construct validity: (a) convergent validity and (b) discriminant validity. *Convergent validity* is usually defined as the correlation between the scores on two or more different instruments on the same construct (Ghiselli, et al., 1981; Green, et al., 1981). Campbell and Fiske (1959) suggested that, for convergent validation, this correlation should be significantly different from zero and sufficiently large. *Discriminant validity* is defined as correlations between tests that should differ from each other. Tests are invalidated if they are too highly correlated with other tests from which they should differ (Campbell & Fiske, 1959; Dickinson & Zellinger, 1980). Campbell and Fiske (1959) suggested two types of evidence for discriminant validation: The correlations between the same constructs on different methods are higher than (a) the correlations between different constructs on different methods, and (b) the correlations between different constructs on the same method.

Discriminability should be differentiated from discriminant validity. This term indicates the ability of rating scales to distinguish those individuals who have more of the construct designed by the items from those who have less of it (between-ratee discriminability). In performance rating scales, it signifies the ability to distinguish the more efficient performers from the less efficient ones. One may figure out discriminability to detect strengths and weaknesses of an individual across many elements in the scales (within-ratee discriminability). Therefore, discriminability can be estimated either by determining the variability among individuals within a dimension (between-ratee discriminability) or by scrutinizing variability among dimensions' ratings within a ratee (within-ratee discriminability) (Jacobs, et al., 1980).

Accuracy

Ratings on scales must be accurate in terms of rating errors or response biases such as (a) halo error, (b) leniency error, (c) central tendency error, (d) logical error, (e) contrast error, (f) similar-to-me error, and (g) proximity error. Even if a given methodology has sufficient reliability and validity, rating errors or biases can inhibit the resulting scores from reflecting the true level of an individual's performance.

Halo Error. A halo error refers to the tendency that a rater appraises individuals in similar ways across traits or dimensions because of the general impression

of the individual. For example, if a child has refused several times to engage in group activity, the teacher may endorse nearly all the 20 items on Self-Adjustment as the child's behavior characteristics, using the Child Behavior Rating Scale (Cassel, 1962). This tendency frequently occurs in a trait that (a) is not easily observable, (b) is not frequently singled out or discussed, (c) is not clearly defined, (d) involves reactions with other people, and (f) is of highly moral importance (Guilford, 1954). It can be avoided by rating one trait at a time on all ratees rather than rating one ratee on all traits at a time, or by using forced-choice formats.

Leniency Error. Raters tend to assess individuals too high or too low, which will result in negatively or positively skewed distributions of ratings. Since a positive leniency (negatively skewed distribution) occurs more frequently, some investigators use a counteracting device as illustrated in Figure 15, in which only one unfavorable descriptive term is given and the other anchors are favorable terms.

Central Tendency Error. Raters tend to hesitate to make positive or negative extreme judgments and want to be safe by rating individuals in the direction of the mean of the group. For example, if a scale includes anchors such as "Never," "Somewhat," "Usually," "Most of the time," and "At all times," a rater may readily select the anchor, "Usually," because it is in the middle position. We can estimate this type of error by calculating the B coefficient (Aiken, 1985). Guilford (1954) assumed that this tendency would be more common in rating those whom the rater does not know well; however, this assumption has not been justified. To avoid this type of bias, participants in the construction of numerical scales should be asked to force the stimuli into a pre-defined distribution (Dawis, 1987). Also the descriptive phrases around the middle should be spaced farther apart in graphic scales (Guilford, 1954). As mentioned earlier, we can avoid excessive use of the middlemost scale point by using an even number of scale points.

Logical Error. An error also occurs when the rater thinks that certain dimensions or traits in a rating instrument are similar, and thus gives ratings in a similar way. For example, if a student usually does not associate with other students, the teacher may rate him/her by circling the number for the item, "Has no friends" on the Walker Problem Behavior Identification Checklist (Walker, 1983). Thinking that the student would naturally feel lonely and unhappy, the teacher may erroneously select the item, "Expresses concern about being lonely, unhappy." This kind of error could be avoided by describing the anchors in operational and observable terms. One of the advantages of Guttman scales and mixed standard scales is their capability to monitor logical errors of ratings and sort out inconsistent raters.

Contrast Error. This type of error takes place when the rater uses him/herself as a referent in evaluating others, and rates them in the opposite direction from

The child's handwriting is:

Poor	Fair	Good	Very Good
1	2	3	4

Figure 15. A scale format for counteracting the positive leniency error.

him/herself in a trait or behavioral dimension. For example, if a rater is usually very tidy in appearance, s/he is liable to assess others to be untidy.

Similar-to-Me Error. This kind of error is somewhat opposite to a contrast error. That is to say, if a rater perceives an individual to be similar to him/herself, s/he tends to rate the individual more favorably.

Proximity Error. A proximity error refers to the tendency to rate a person on adjacent traits or dimensions in a similar fashion. For example, if items on Social Adjustment are immediately followed by items on School Adjustment, the rater may respond to the two sets of items in almost the same way. This tendency may be avoided by placing similar traits or dimensions farther apart and more obviously different ones close together, or by rating one trait at a time with greater time intervals between them (Guilford, 1954).

On the whole, these errors can be minimized by training raters to be properly acquainted with them and to critique example ratings on the same type of scales as those to be used for actual rating, in terms of such errors (Blanz & Ghiselli, 1972; Green et al., 1981; Guilford, 1954).

Qualitative Criteria

Qualitative criteria are those rules or guidelines by which we evaluate performance rating scales on their adequacy, usefulness, and benefit. These criteria include (a) relevancy, (b) data availability, (c) practicality, (d) equivalence, and (e) interpretability.

Relevancy

Behaviors specified in the rating scales must be important to the successful performance of a task. Also, these behaviors must exclude those which are not related to the task. These requirements can be met by an "in-depth job analysis" (Jacobs et al., 1980, p. 606). Or, professionals who have sufficient knowledge of, and contact with the individuals to be rated may scrutinize the items and dimensions of the scales for their relevancy. Quantitative criteria of relevancy are referred to as content validity.

Data Availability

This qualitative issue refers to the availability of information on three questions: (a) Is direct observation possible? (b) Who should do the evaluating? and (c) What additional data, relevant to performance, are available? If direct observation that is related to a given construct is not possible, ratings by informants probably cannot be used. Rather, individuals should rate their own behaviors using self-rating scales. Additional information from other sources may support or validate the ratings.

Practicality

Issues of time, cost and efforts for both developing and administering rating scales must be taken into consideration. For example, if a set of BARS and a set of short-cut BARS have the same level of psychometric properties, the short-cut BARS should be used. Instructions on administering a rating scale should be succinct and clear. The scoring and reporting system should be simple.

Equivalence

When two or more raters evaluate individuals engaged in similar or identical performance tasks, equivalent frames of reference or common standards and objectives should exist. Thus the ways in which different raters evaluate individuals' performance behaviors under a specific situation should be comparable. For example, if ratings of some handicapped children on a specific social skill are to be compared with the performance of their normal peers, the comparison should be based on comparable settings and performance tasks.

Interpretability

The conditions should exist that all raters, or even ratees, evaluate observed behaviors in a similar way with regard to the behavioral examples or anchors on the scales. They should also perceive the documented examples in a similar way. To achieve uniformity in interpretation, the documented examples must be stated in observable and descriptive terms rather than evaluative or inferential terms. Raters should also possess sufficient knowledge about, or familiarity with, the performance task.

Utilization Criteria

Utilization refers to the purpose for which a rating is conducted. Attitude scales can be used, for example, as references for grouping/organization, selection, performance strategies, new policies, or behavior modification. In addition, this type of self-rating can provide indirect and supplementary information to informant (performance) ratings.

Utilization of rating scales on overt behaviors (or performance) is usually evaluated against some criteria as suggested by Jacobs et al. (1980), from which behaviors that are considered relevant to educational settings have been selected: (a) disciplinary action, (b) feedback, (c) promotion, (d) selection/validation, and (e) training/supervision. Establishment of a purpose prior to a rating helps investigate whether or not the results are used appropriately (i.e., content or criterion-related validity).

Disciplinary Action

Ratings on scales can be used to identify individuals whose performance is "less than satisfactory" to demote or refer them to lower levels for more training or practice, or to give them warnings or dismissal. For this purpose, behavioral rating scales should have sufficient "less than satisfactory" items to monitor more minute and specific behavioral deficiencies.

Promotion/Awarding

Rating scales can be used to distinguish those who are "above average," to promote them to higher levels or to award them. Promotion or awarding based on this rating presupposes that the present behavior characteristics are requisites for the future position. Therefore, only those overlapping behavior dimensions between the present and the future positions should be considered (Jacobs et al., 1980).

Feedback

Sometimes performance rating scales should provide the individuals being examined with sufficient, concrete, specific information on their present level and on what is needed. This information is especially important when the purpose of the rating is discipline, or the behavior being rated is a prerequisite for another behavior. Such information is most beneficial because it can decrease the likelihood of a disciplinary action, and increase the likelihood of being awarded by enhancing or improving their behaviors. It also helps make behavioral objectives become explicit.

Selection/Validation

This criterion refers to the ability to generate performance scores used in a regression equation—an equation for a straight line through the means in a scatter diagram, which is usually designated as "$Y = a + bX$" where a is the Y intercept and b is the regression coefficient. For example, suppose that X represents the age and Y represents the score on a rating scale, and the distribution of scores across age levels is linear for a certain population. Then the above equation best predicts the average score at a specific age level. Therefore, the regression line represented by the equation divides individuals approximately as falling above and below the average score at the age level.

One primary function of this evaluation is to select individuals for some purpose, such as identifying those individuals with deficient behavior for disciplinary action, or those with excelling behavior for promotion/awarding. Another function of this evaluation is to

assess the validity of a program. Any performance evaluation methodology should have this function.

Training/Supervision

Training and program objectives can be established by the information obtained through a scale construction procedure. Or, the dimensions in a set of rating scales can serve as criteria for training or supervision. Thus, an effective performance rating scale can benefit both the trainees and the trainers because (a) The trainees can set goals of performance based on the criteria, or by a model of high performers whom the evaluations indicate; and (b) Trainers, teachers, or supervisors can obtain information on their students' or subordinates' current levels of performance and guidelines to improve their skills and to reinforce high performers.

Section Summary

Behavior rating scales, whether they are designed to assess attitudes or to measure performance levels, must be evaluated by some generally accepted criteria so we can trust the results of ratings with those scales. Quantitative criteria are related to how accurately and consistently rating scales measure those qualities represented by a behavior domain, criterion, or construct. Qualitative criteria provide guidelines to ascertain whether specific scales include critical components for a task or a construct, and how effectively and efficiently they are used. Jacobs et al. (1980) reported that the qualitative criteria have been ignored by any performance evaluation system. Utilitarian criteria act as a check to investigate what purposes certain scales serve. In addition to these criteria, Rie and Friedman (1978) suggested some guidelines on item configurations to enhance the value of scales: (a) Items should be defined in operational terms so that they denote the same meaning to all raters; and (b) Items must be descriptive, but not be evaluative or inferential in nature. These guidelines should be used to reduce the rating errors mentioned earlier.

Published Rating Scales

Numerous behavior rating scales have been published to assess behavior problems in school-age children. This section provides teachers and researchers with guidelines for selecting published rating scales, and with a brief overview of some of those rating scales. For more comprehensive review of rating scales and checklists, readers may refer to Rie and Friedman (1978), Edelbrock (1988), Witt, Cavell, Heffer, Carey, and Martens (1988), and Wilson and Bullock (1989).

Informant-Rating Scales

We sometimes evaluate a child's behaviors by using a series of informant-rating scales, which require an observer and rater other than the child him/herself to be rated. Informant-rating scales are appropriate when teachers, parents, or other related persons are able to observe the child's overt behaviors that are related to a defined behavior domain or construct. They have the following advantages and disadvantages.

Advantages

1. We can report the results of the rating in more objective, observable terms. This objectivity permits us to make more direct comparisons among children to determine their similarities and differences in reported behaviors.

2. Informant ratings report recent or current behaviors because the observation period usually comes just prior to the rating.

3. More published rating scales are available for the informant rating than for the self-rating.

Disadvantages

1. Different raters may have different types and amounts of exposure to the target child, and this difference may cause their ratings to be biased (Edelbrock, 1988). This problem can be minimized by setting a specific observation period for all the raters.

2. The characteristics of informants can influence their ratings in such a way that they cause rating errors described in chapter 2 (Edelbrock, 1988). These errors can be prevented to some degree by sufficient rater training.

3. One impressive incident during the observation may cause the rater to overemphasize a certain aspect of a child's behavior. For example, if a child broke a window in the classroom by mistake, this incident may cause the teacher to rate the child as "clumsy."

Guidelines for Using Informant-Rating Scales

Although the manual of any rating instrument provides details on the behavior domain, standardization data, and administering and scoring instructions, we suggest two additional considerations on using the instrument: (a) a sufficient observation period, and (b) rater training. These two considerations are critical to prevent errors and bias in rating.

Self-Rating Scales

Some rating scales are designed so that individuals rate their own behaviors. This type of rating instrument is appropriate when the target traits are covert or abstract, such as attitudes and self-concept. Informant-rating scales cannot serve appropriately because these traits or constructs cannot be described in observable terms. Self-rating scales have the following advantages and disadvantages.

Advantages

1. Through self-ratings we can obtain valuable information not available through other methods of assessment because the information thus obtained is related to what the child has to say as well as how he or she says it.

2. Children's responses to those items about treatment preferences can be used to design interventions that are more acceptable to children.

3. We can get information about reinforcement preferences and peer interactions through ratings on self-rating scales (Witt et al., 1988).

Disadvantages

1. Children may fail to understand the contents described in items because of their vocabulary or grammar levels.

2. The rater or ratee tends to remember and distort information in systematic ways as a function of age and cognitive ability.

3. The rater or ratee tends to oversimplify and to interpret what s/he sees in an all-or-nothing manner, which can bias self-rating. For example, young children may perceive a mother as someone who is cooking and a father as someone who mows the lawn. To them, if a mother mowed the lawn, she could not be a mother. Likewise, if a father cooked food, he could not be a father (Witt et al., 1988).

4. Children may not respond truthfully to those items that include socially or morally sensitive issues such as drugs and sex.

Guidelines for Modification

Since the advantages described above seriously threaten reliability and validity of the ratings, Witt et al. (1988) suggest some guidelines for modification of self-rating scales:

1. Stimulus complexity should be adjusted to the child's cognitive and language facility.

2. The response format should be simplified by (a) providing the children with a limited number of predefined oral response options to an item to select the best response, (b) allowing the child to respond to a limited number of predefined written options by choosing the best response, or (c) allowing the child to respond by selecting a concrete representation of his/her response, such as one picture or object.

3. Prior to evaluation, children should be fully trained on the expectations of what the rating is about and how they should respond.

Summary

Based on the reviews by Edelbrock (1988) and Witt et al. (1988), we suggest the following considerations for selecting a rating scale either for rating by the informant or for self-rating:

1. Read the manual for the rating instrument to find what type of behavior and what domain of the behavior it is designed to measure. Some rating scales, e.g., *The Burks' Behavior Rating Scales* (Burks, 1977), focus on global attributes, whereas others, e.g. *The Piers-Harris Children's Self-Concept Scale* (Piers & Harris, 1969), are designed to measure smaller, or more specified ranges of behavior.

2. Investigate the extant psychometric properties such as reliability and validity.

3. Ensure that the rating scale is appropriate for the ratee's and/or informant's age, cognitive, or educational level, and language capacity. To be certain of the contents, review the individual items.

4. Examine the response format to determine the difficulty and/or burden of the measure. It should be simple to use to avoid any errors in rating. Also the number of items should not be too large, considering the time and effort needed for rating and scoring.

5. The financial costs of the assessment materials should be taken into consideration.

In the following section, we will review some behavior rating instruments.

Overview of Measures

The Walker Problem Behavior Identification Checklist

Scale Features and Scoring

The Walker Problem Behavior Identification Checklist (WPBIC) developed by Walker (1983) has two different forms of checklist—one for females and one for males. Either form contains five behavior dimensions, each of which represents one scale: (a) Acting-Out, (b) Withdrawal, (c) Distractability, (d) Disturbed Peer Relations, and (e) Immaturity. It also includes 50 items (statements), each of which belongs to one of the five scales. All these items describe negative behaviors. The score weight for each item ranges from one to four. The scales are vertically arranged graphic scales; that is, the anchors (statements) for each scale are arranged from top to bottom with scale points assigned to each.

The rater should be a classroom teacher—the individual who spends much more time than any other school personnel actually observing the child. Walker (1983) recommended that the WPBIC be used to rate the child 2 months after school starts, to allow for an observation period. When using the WPBIC, the rater is asked to circle the number in one of the five columns to the right of each statement, if s/he observed the behavior described in that statement during the observation period. If s/he did not observe that behavior, the rater should skip that item without making any mark on it. Thus the scales basically have a checked/unchecked format.

When rating on the individual items, it is not important to know which item belongs to which scale or behavior dimension, because the items are arranged in random order across dimensions without any labels for the dimensions. When the rating is completed, the rater adds up the values of the numbers in each column and

writes the sum in the square at the bottom of the column. Then s/he adds up all the scores in the squares across columns to calculate the total score. These column scores and the total score are plotted on a Profile Analysis Chart, which shows their relative standings and their corresponding *T*-scores ($M = 50, SD = 10$).

Norms

The normative procedures were originally performed with a sample of 21 teachers from the pool of 4th, 5th and 6th grade teachers in a local school district in Oregon. They rated 534 regular school children in the fourth, fifth, and sixth grades on the WPBIC. This procedure yielded a mean raw total score of 7.76 with a standard deviation (*SD*) of 10.53 (Walker, 1976).

Subsequently, the normative procedures included (a) a preschool/kindergarten sample with 29 teachers who rated 469 children ages 2 to 6 ($M = 10.3, SD = 8.76$), (b) a primary sample including 35 teachers who rated 852 children of grades 1, 2, and 3 ($M = 5.61, SD = 9.37$), and (c) a handicapped sample including 40 severely learning disabled, mentally retarded, behaviorally disturbed, and communication disordered children who were rated by their regular and/or special education teachers ($M = 16.93, SD = 15.56$).

Reliability

The reliability of the WPBIC was examined by a split-half method. Items were arranged in such a way that the score weights for the first half of the items were equal to those for the second half. For example, if an item with a score weight 4 is selected as Item 1, another item with a weight of 4 would be selected as Item 50. The remaining 48 items were also selected in this way. The split-half reliability coefficient was .98 with a standard deviation of 10.53 and a standard error of measurement of 1.28. This coefficient is well above the minimum acceptable coefficient of .90 for making individual discriminations among subjects.

The test-retest reliability of this instrument was also investigated by two studies. In the first study, 200 children in grades 1 through 6 were rated on the WPBIC by their teachers two times within a period of 3 weeks. The overall coefficient was .80 for a 3-week interval, and for individual teachers, the coefficients ranged from .43 to .96. The second study employed this instrument to rate two groups (samples) of students in the third and fourth grades. The coefficients were .89 for the first sample (N=30), .81 for the second sample (n=36) and .86 for both.

Validity

Content Validity

Thirty experienced teachers were asked to give operational descriptions of their pupils' behavior problems in the school or classroom setting. The interviews with the teachers yielded an item pool of 300 items, from which 50 most frequently mentioned behaviors were selected for inclusion in this instrument. This procedure has similarity with the critical incident technique mentioned earlier. Therefore, this checklist has strong content validity and maximal relevance for use by teachers or other school personnel.

Criterion Validity

The scores of the students in the normative samples were compared to several criteria. First, 38 students from the norm sample of behaviorally disturbed children were compared to the same number of students who were not identified as being behaviorally disturbed, in terms of their scores on the WPBIC. The two groups were matched in terms of age, grade and sex. The difference of means between the two groups was significant ($D = 10.16$, $CR = 4.23$, $p<.001$). Thus contrasted-groups validity was considered to be appropriate.

Second, teacher rankings of child interactive frequency and ratings on the Withdrawal scale of the WPBIC were compared to social interaction rates, which were recorded by direct observation in free play settings. Intercorrelations between teacher rankings of interaction and ratings on the WPBIC ranged from -.478 to -.615 across three time points ($p<.01$). Stronger correlations were obtained between these two and the criterion variable of social interaction rates.

Third, two studies investigated the relationship between academic achievement and WPBIC total scores. The resulting correlations were -.32 and -.34, respectively. Therefore, higher achievement scores were associated with lower WPBIC scores as was expected by Walker (1983).

Fourth, teacher ratings on the WPBIC were compared to the direct observation data on the three most and least deviant children in terms of their appropriate classroom behavior. The average scores of the three most deviant children and the three least deviant children on the WPBIC were 18.00 and 1.00, respectively.

Fifth, there was a highly significant correlation between teacher ratings and parent ratings on the WPBIC ($r = .81, p<.01$). This reflects consistency across settings and across raters.

Sixth, 12 deviant and 12 nondeviant boys (ages 6 to 11) were rated on the WPBIC by their parents. The deviant group received an average score of 47, with a range of 29 to 67. The mean score for the nondeviant group was 12, with a range of 1 to 24.

Construct Validity

Six studies demonstrated evidence for the construct validity of the WPBIC by showing its sensitiveness to empirically documented behavioral changes. Five of them employed the WPBIC as one

dependent measure to examine the effects of (a) systematic intervention on problem behaviors, (b) a crises-resource program, (c) teacher intervention on social withdrawal, (d) interventions on conduct disorders, and (e) parent training on child behavior ratings. For example, in one of the studies, each of the 50 regular teachers rated a single target handicapped child on the WPBIC before (i.e. pre-test) and after (i.e. post-test) the intervention. The average number of problem behaviors on the post-test for the experimental group was 7.9, while that for the control group was 10.6. The difference between these two means was statistically significant ($p<.05$).

The remaining study also utilized the WPBIC as a dependent measure. Parents rated their adolescent children who had been referred to a clinic for diagnostic consultations and short-term counseling. The mean total score of the group was 22.24 (SD = 1.49), which indicated that the children were highly deviant.

The Behavior Rating Profile-Second Edition

Scale Features and Scoring

Unlike other scales, the Behavior Rating Profile-Second Edition (BRP-2) consists of six instruments (Brown & Hammill, 1990). It includes three self-rating scales, two informant-rating scales, and a sociogram, which is excluded in the following summary because it is not a rating scale or checklist but a peer nomination technique. The self-rating scales (all student ratings), in turn, embrace three instruments: (a) Home, (b) School, and (c) Peer. Among these, the first two solicit the student's self-report about his/her behavior in two ecological settings (i.e. home and school), and the Peer scales solicit the student's self-reported information on his/her social skills or interpersonal relationships. The two informant scales (i.e. the Teacher Rating Scale and Parent Rating Scale) provide information about the student's behavior from different sources. Through these multiple measures, we can obtain more accurate, multidimensional information about a child's behavior. This multidimensionality agrees with Fish's (1988) suggestion that reality contains multiple variables and that, therefore, multivariate methods must be used to study it.

The Teacher Rating Scale and the Parent Rating Scale have 30 items each, with four anchor points: (a) "Very Much Like the Student (My Child)," (b) "Like the Student (My Child)," (c) "Not Much Like the Student (My Child)," and (d) "Not At All Like the Student (My Child)." The three Student Rating Scales contain 20 items each, with all 60 items combined into a single instrument, presented in random order. Students are instructed to respond to each item by checking "True," if they believe the item describes themselves well, or "False," if the item does not describe themselves.

The student self-rating scales (i.e., Home, School, and Peer) and the informant-rating scales (i.e., the Teacher Rating Scales, and the Parent Rating Scales) differ from each other in terms of computing raw scores from the protocol. For the student rating scales, the teacher counts the number of False responses for each scale and finds the appropriate column and row in the table provided in the manual to convert it to a standard score (M = 10, SD = 3) or percentile rank. For the informant-rating scales, the anchors are weighted: 0 for Very Much Like the Student (My Child), 1 for Like the Student (My Child), 2 for Not Much Like the Student (My Child), and 3 for Not At All Like the Student (My Child). The examiner counts the number of responses for each of these anchors and multiplies it by the corresponding weight. The raw score is the sum of these products across the anchors. This score is also converted to a standard score or percentile rank by referring to the appropriate table in the manual.

Norms

The participants in the standardization procedures for the BRP-2 Student Rating Scales were 2,682 children living in 26 states in the United States. Their ages ranged from 6.6 to 18.6 years, but their grade levels and the proportion of boys and girls were not specified in the manual. This sample did not include any children "formally identified by a multidisciplinary evaluation team as Severely Emotionally Disturbed" (Brown & Hammill, 1990, p. 39).

The normative sample for the BRP-2 Parent Rating Scale included 1,948 parents from 19 states. They were asked to rate any of their children "who had not been formally identified by a multidisciplinary team as Severely Emotionally Disturbed" (Brown & Hammill, 1990, p. 39). Whether or not these children overlap those in the sample for the Student Rating Scales is not specified.

The participants in the normative procedure for the BRP-2 Teacher Rating Scale were 1,452 classroom teachers who were teaching in 26 states. They were randomly selected and then every fifth student on their alphabetically listed class rosters was rated. Only those students who had been "formally identified by a multidisciplinary team as Severely Emotionally Disturbed" (Brown & Hammill, 1990, p. 39) were excluded from the normative sample. Again, whether or not those children included overlap those in the sample for the Student Rating Scales is not specified.

Reliability

The reliability of the BRP-2 was estimated by (a) internal consistency, (b) test-retest reliability, and (c) standard error of measurement. These are described below.

Internal Consistency

To investigate whether or not the items in each of the BRP-2 scales are intercorrelated (i.e. homogeneous) and thus measure the same construct, Cronbach's (1951) coefficient Alpha was used. This procedure produces "the average of all possible split-half correlations that can be extracted from a test" (Brown & Hammill, 1990, p. 44). Examples were drawn from the normative samples, and coefficients Alpha were calculated by using their protocols within five different grade intervals: grades 2-3, grades 4-5, grades 6-7, grades 8-9, and grades 10-12. The resulting coefficients Alpha showed that only 3 of the 25 coefficients did not meet or exceed the .80 criterion.

Test-Retest Reliability

Brown and Hammill (1990) referred to two studies pertaining to the test-retest reliability of the BRP-2. In one study, the reliability was examined with 36 Indiana high school students, 27 of their parents, and 36 of their teachers, allowing a 2-week interval between the two administrations. The ensuing coefficients ranged from .78 to .91, with only one of them dropping below the .80 limitation. In the other study, the test-retest reliability was investigated with 198 students (55% males and 45% females), 212 parents, and 176 teachers in central Michigan, permitting an interval of 14 to 16 days. The subjects were grouped into 2-year grade intervals (e.g. grades 1 and 2, grades 3 and 4, and so forth). The scores on the pre-test and post-test were correlated; the resulting coefficient ranges are as follows: .43 to .91 for the Student Rating Scales: Home, .58 to .92 for the Student Rating Scales: School, .52 to .90 for the Student Rating Scales: Peer, .69 to .96 for the Parent Rating Scale, .90 to .96 for the Teacher Rating Scale.

Standard Error of Measurement. Standard error of measurement was calculated by taking the square root of 1 - r (i.e., each coefficient Alpha), multiplied by the standard deviation. This value was used to show the extent of deviation from the mean of the standard scores due to error for each score. All the BRP-2 scales had small standard errors of measurement ranging from 1 to 1.5 points, which suggests high reliabilities.

Validity

Content validity was established by (a) content validation procedure, and (b) empirical item selection procedure. For the content validation procedure, the constructs or characteristics that would be measured were clearly defined. Then, items were obtained from two sources: (a) professional literature, and (b) significant informants. The significant informants included parents and teachers of emotionally disturbed and learning-disabled children. These items constitute item pools for the experimental version.

The empirical item selection was based on two criteria: (a) item discrimination, and (b) item difficulty. After administering all the items in the experimental version, item discrimination coefficients were calculated by correlating the scores on the items to the total score to assure each item's significant contribution to the instrument. Those items were selected whose discrimination coefficients were significant at or beyond the .05 level, and fell between .30 and .80. The other criterion, item difficulty, is defined as the percentage of test subjects who give a correct answer to the item. For BRP-2, items describe negative behaviors, and therefore their presence or observation indicates behavioral problems. Anastasi (1988) asserts that a behavior that is observed with significant frequency in a substantially representative sample cannot be an indicator of abnormality. Therefore, those items were selected whose median percentages of unfavorable responses fell below the 50% mark.

Discriminant validity was addressed by correlating the scores on the BRP-2 scales to those on measures of achievement, and on measures of aptitude. The results indicated that no significant relationship existed between the BRP-2 scales and any of the criterion measures.

Convergent validity was also addressed by correlating the scores on the BRP-2 scales and those on eight measures of behavior or effect. The results showed that the BRP-2 scales are highly correlated with those criterion measures.

Construct validity was inspected by examining the intercorrelations among the subtests of the BRP-2. The resulting 40 coefficients were significantly high, ranging from .49 to .96.

Burks' Behavior Rating Scales

Scale Features and Scoring

Burks' Behavior Rating Scales (BBRS) were designed to identify particular problems or patterns of problems of behavior shown by children in grades 1 through 9 who are referred to school or community counselling agencies for behavior difficulties (Burks, 1971). Therefore, this instrument is not appropriate for screening groups of normal children. It measures the child's overt or observable behaviors, not the "inner world" (Burks, 1977, p. 5). The raters or informants should be those persons who are well acquainted with the children, or who can obtain information about the children.

The BBRS contains 110 items in nine behavior categories. The scales are presented in a 5-point Likert-type format: The rater is required to respond to each item by selecting one from the five alternatives or anchors ranging from, "You have not noticed this behavior at all" (1 point), to, "You have noticed the behavior to a very large degree" (5 points). Thus, the rater enters the appropriate number in the box pro-

vided on the right-hand side of each item. For scoring, the rater adds together the numbers in the five boxes for each category and transfers the total to the profile sheet.

Norms

Regular-Class Students

Four hundred ninty-four primary age children (grades 1 through 3) and 69 older children (grades 7 and 8) in regular classes were evaluated on the BBRS by their teachers. The results illustrated the following two issues (Burks, 1977, p. 31):

1. The majority of school children did not show the pathological symptoms described in the scales. That is, most ratings fell in the "not significant" classification.
2. The percentages of ratings falling in the three classifications ("Not Significant", "Significant", and "Very Significant") varied considerably from one behavioral category to another.

Disturbed Children

The BBRS was used to assess the behavior characteristics of 267 children who were later placed in educationally handicapped classes or were given prescriptive diagnostic help in regular classrooms. These children included 153 primary age (grades 1 through 3) children and 114 elementary age (grades 4 through 6) children. Their scores demonstrated that significant distribution differences existed between this group of referred students and the group of regular class students. In the group of referred students, boys outnumbered girls (114 boys versus 39 girls at the primary level, and 91 boys versus 23 girls at the elementary level). However, the referred girls showed many more difficulties than average-class girls, and their behavior pattern was the same as that of the "disturbed" boys in six categories: (a) poor academics, (b) poor attention, (c) poor ego strength, (d) poor coordination, (e) poor intellectuality, and (f) excessive withdrawal.

Reliability

Item reliability was examined by rating and rerating 95 "disturbed children" within a period of 10 days. Ratings on "normally behaved children" had very high item reliability. Raters gave a great majority of the children the 1 rating (i.e., You have not noticed this behavior at all) on most items, and they rated them the same at a second time. All items demonstrated high correlation coefficients, ranging from .60 to .83 with the average of .705.

Validity

Content Validity

The items and categories of the BBRS were conceptualized, modified, and repatterned over a 4-year period. Twenty-two school psychologists in Los Angeles County, California, acted as judges to examine those items and categories; their suggestions contributed to the improvement of the content validity.

Factorial Validity

Factor-analytic studies of the ratings on the BBRS pertaining to age factors indicated (Burks, 1977, p.36):

1. The younger the age group, the greater the number of discovered statistical factors.
2. The younger the age group, the fewer the number of scale categories found to be included in each factor.
3. The older the age group, the less clear the differentiation between factors.
4. An aggressive acting-out factor occurred at all age groups.
5. An immature factor was found at the primary age level that appeared at older age levels.
6. Neurotic factors change dimensions and seemingly have different meanings from one age level to another.

Construct Validity

One hundred seventy-six children rated themselves on the School Attitude Survey (Burks, 1970). Based upon this self-rating, 25 students who reported the most inner disturbance and 25 students who reported the least inner disturbance were selected for rating on the BBRS, which measures the outward behaviors of children. Teachers rated these children on the BBRS. A highly significant relationship was found between the teacher ratings on the BBRS and the self-ratings on the School Attitude Survey.

Contrasted-Groups Validity

This type of validity is associated with the ability of the instrument to differentiate between two independent groups that are related to the definition of the construct being measured. It was estimated by correlating scores of the two groups (i.e., the cross sample group from regular classrooms and the group of the referred students). As mentioned in the previous section, the referred group had higher category ratings than the cross-sample group.

The Child Behavior Rating Scale

Scale Features and Scoring

The Child Behavior Rating Scale (CBRS) developed by Cassel (1962) contains 78 items which are classified into the following five categories: (a) Self Adjustment, (b) Home Adjustment, (c) Social Adjustment, (d) School Adjustment, and (e) Physical Adjustment. Since this instrument is an informant rating scale, it must be used "only by raters who have observed or know directly the behavior of the child to be rated" (p. 1). The CBRS can be used for the following purposes:

1. Achieve objectivity in ratings of the behavior of children by raters who have observed or known those children.
2. Compare ratings of a child with the normative

data of both normal ("typical") children and emotionally handicapped children.

3. Provide objective measurements in five adjustment behavior dimensions.

4. Provide a single meaningful score as an indicator of the Personality Total Adjustment Score (PTAS).

5. Obtain objective comparisons between ratings of the same child made by different raters.

6. Help understand the interpersonal relationships between different raters (mother, father, teacher, etc.) and the child by comparing ratings by those raters.

7. Help understand the dynamics of the home (mother and child, father and child, etc.) by comparing ratings of different raters.

8. Facilitate research studies of the young child, especially the young child in his first years of adjustment to the school situation.

Response choices for each item are presented in a 6-point Likert-type end-defined format (Dixon, Bobo, & Stevick, 1984). That is, the configuration of the response choices looks like the following:

Yes 1 2 3 4 5 6 No

Under each of these scale points, a column of boxes is provided, in which the rater makes marks if s/he chooses that point on the items. Steps for scoring are:

1. Count the check marks in column boxes within a behavior dimension and put this number in the box for "Number Checks."

2. Multiply the number of marks by the number of the column, the product of which is the *weighted value* for that column. For example, if the rater has made 10 marks in the No. 3 column boxes for 10 items within the Self-Adjustment area, the weighted value is 10 x 3 = 30.

3. To compute the Personality Total Adjustment Score (PTAS), use three of the five adjustment-area weighted scores. For this score, (a) Multiply the Self-Adjustment weighted score by 2, (b) Multiply the Home Adjustment weighted score by 2, and (c) Add the School Adjustment weighted score to the sum of these scores. The sum total is the PTAS.

Norms

Two normative groups were rated on the CBRS: a group of 2,000 typical (normal) children, and a group of 200 maladjusted children. The children in both groups were in the preschool and primary grades. The weighted scores of the five adjustment areas, and the PTAS, were converted into McCally *T*-Scores. The standard error of the McCally *T*-Score for the CBRS was computed. The standard error of the normal group data was 4.88 *T*-Score points, and that of the emotionally handicapped group data was 8.00 *T*-score points. Using the McCally *T*-Score, we can compare a single child with the two normative groups.

Reliability

Using the split-half method, indices of internal reliability were computed for the entire CBRS on a sample of 800 typical children, and on another sample of 200 maladjusted children. The resulting Pearson Correlation Coefficient (r) for the typical group was .873 + .003 and the r for the maladjusted group was .589 + .042.

By the test-retest method, inter-rater reliability was estimated. The r on a sample of 50 parents was .913 + .024, and the r on a sample of 50 teachers was .739 + .065.

The five adjustment area scores on the CBRS were inter-correlated. The results indicated that moderate r's were obtained for the three groups (teachers, mothers, fathers), and that, of the five adjustment-area scores, the Physical Adjustment scores were least related to the other four area scores.

Validity

Construct Validity

Scores on the CBRS are highly correlated to scores on other psychological instruments, such as school achievement test scores, intelligence quotients, and social development. Those correlation data indicated that the CBRS is an effective instrument for predicting performances on the Metropolitan Achievement Tests, I.Q., and social quotients of the Vineland Social Maturity Scale. The Pearson Correlation Coefficient between ratings of mothers and fathers was .656 + .023, which is high enough for statistical and clinical significance. Therefore, either parent's ratings can be used as an estimation of adjustment behaviors, instead of having ratings of both parents.

Status Validity (Discriminability)

A sample of 200 randomly selected normal children ranging 5 to 9 years (M = 6.2 years, SD = 0.6 years) was compared with a group of 200 maladjusted children from the same community in the same age range (M = 6.8 years, SD = 0.8 years). This latter group consisted of children referred for psychological services because of behavior adjustment problems. Ratings of the two groups on the CBRS were significantly different, and two other studies of status validity yielded similar results.

The Behavior Evaluation Scale-2

Scale Features and Scoring

The Behavior Evaluation Scale-2 (BES-2), from which the present version originated, included 50 items (McCarney, Leigh, & Cornbleet, 1983). These items were classified into five subscales representing the characteristics of the term "seriously emotionally disturbed," which are specified in the PL 94-142 (Education of All Handicapped Children Act) (Federal Regis-

ter, 1977, p. 42478). The authors titled the subscales as follows: (a) Learning Problems, (b) Interpersonal Difficulties, (c) Inappropriate Behaviors, (d) Unhappiness/Depression, and (e) Physical Symptoms/Fears.

The Behavior Evaluation Scale-2 (BES-2) was designed primarily to identify behavior disorders and emotional disturbance of school children in grades K-12 (McCarney & Leigh, 1990). The authors suggested six major purposes for which the BES-2 can be used:

1. Pre-referral screening and identification of problem behaviors.
2. Comprehensive behavioral assessment for post-referral procedure.
3. Diagnosis of behavior disorders and emotional disturbance for determining eligibility for special services on legal criteria.
4. Provision of specific information about "strengths and deficits" (p. 3) in individual students' behavior, which can assist in development of Individualized Education Programs.
5. Documentation of behavioral progress made by individual students.
6. Collection of objective and quantifiable data regarding the frequency and severity of behaviors in question. (McCarney & Leigh, 1990, pp. 2-3)

The BES-2 comprises 76 items depicting specific observable and measurable behaviors. The items are divided into the same five subscales as specified in the BES.

The PL 94-142 definition of behavior disorders and emotional disturbance relates that the specified traits must occur "to a marked degree" for eligibility of special services in that area (Federal Register, 1977, p. 42478). McCarney and Leigh (1990) interpreted the phrase "to a marked degree" as indicating "both frequency and severity of behaviors" (p. 5). Aligned to this notion, the seven anchors for each item describe specific, objective frequencies of a behavior.

These descriptors of frequencies indicate more accurate calibration of the scale than typical subjective descriptors, such as "frequently" or "sometimes."

While either the items or the descriptors of them are weighted in most published rating scales, both the items and the descriptors in the BES-2 have weights according to the seriousness and the frequency of the specified behaviors. This dual weight system purports to enhance the construct validity of the instrument by measuring the two behavioral dimensions. The Data Collection Form, which is a supplemental form designed to be used to gather frequency data from direct observation, can further contribute to the accuracy of measurement.

Due to the dual weighing system, the scoring procedure is slightly more complicated than those of other scales reviewed in this section. The steps are as follows:

1. Multiply the rating score for each item by its weight.
2. Calculate the raw score for each subscale by adding the weighted scores for the items within the subscale.
3. Convert the subscale raw scores to standard scores.
4. With the sum of the standard scores, locate the Behavior Quotient, the standard error of measurement (SEm), and the percentile rank in the specified tables in the manual.

Norms

The normative sample included 2,272 students from 31 states, who were "fairly evenly" (p. 9) distributed across grade levels K-12. This distribution was "intentionally designed to ensure appropriateness of the norms for all grade levels" (p. 9). The BES-2 was administered to randomly selected students from 568 regular classrooms. The normalization procedure was conducted from the Fall of 1988 through the Spring of 1989.

Reliability

Internal Consistency

To estimate the extent to which the items in each scale indeed measure the same construct (i.e., internal consistency), the coefficient Alpha procedure (Cronbach, 1951) was employed. The result showed that sixteen of the twenty-four coefficients reached or exceeded .90, six of them exceeded .80, and the remaining two were .75 and .78 respectively. All of them were statistically significant at the .05 level. Thus the result demonstrated strong evidence of internal consistency.

Test-Retest Reliability

To examine the stability of BES-2 over time, the instrument was administered to 82 "normally-achieving" (McCarney & Leigh, 1990, p. 12) students, and 108 students who had been diagnosed as behaviorally disordered (BD). These students were divided into two parts, and rated by their teachers on the BES. With the ensuing data, Spearman correlation coefficients between the two sets of scores from the two parts of students were calculated for each of the five subscales and for the total scale. All the coefficients except one exceeded .90, with the remaining one of .89. McCarney and Leigh (1990) attributed this stability to the objective nature of the descriptors, and the clear descriptions of behaviors.

Validity

Content Validity

Content validity of BES-2 was established from two sources: (a) derivation of items, and (b) review by teachers and professionals. The original item pool of

the Behavior Evaluation Scale (McCarney, Leigh, & Cornbleet, 1983) was developed by teachers of behaviorally disordered students in Missouri. With 47 items selected from the pool, teachers and professionals were asked to eliminate inappropriate ones, to modify items with unclear wording, and to add appropriate ones. Further modifications were made through field testing by elementary and secondary level classroom teachers.

In 1988, 31 new items were added to the BES to form a new item pool of 83 items for BES-2. These new items were suggested by teachers and professionals in the area of behavior disorders/emotional disturbance. A group of 675 teachers judged the appropriateness of the items, eliminating inappropriate or unimportant items, adding new ones, and modifying existing ones. All the 76 items selected for the final scale were those that at least 95% of the teachers approved as being appropriate.

Convergent Validity

To examine the degree to which the BES-2 correlates with another measure of the same construct, scores on the BES-2 were correlated with scores on the Teacher Rating Scale of the Behavior Rating Profile (BRP) (Brown & Hammill, 1978). The correlation between the total scale scores on the BES-2 and those on the BRP ($r = .76$) was statistically significant at the .01 level. Among the five correlation coefficients for the five subscales, the coefficients for Interpersonal Difficulties ($r = .73$), Inappropriate Behaviors ($r = .81$), and Physical Symptoms/Fears ($r = .62$) were statistically significant at the .01 level, and the coefficient for Unhappiness/Depression was significant at the .05 level. The only subscale that did not reach a statistically significant coefficient was Learning Problems.

Criterion-Related Validity

A study was conducted with 190 students and the same number of regular and special education teachers to compare the results on the BES-2 with the teachers' professional judgment about the behavior of the students in the group that was composed of normal and BD students. The teachers first responded to a question on their students' status in classroom behavior relative to other students of the same age, by selecting one of the descriptors on a 9-point scale. Next, they rated the same students on the BES-2. All the correlation coefficients between the results on the BES and the criterion (i.e., teachers' judgment) were statistically significant at the .01 level except the one for the Physical Symptoms/Fears subscale for the BD students. One threat to the validity of the ratings on the BES is the possibility of a halo effect, because the teachers can acquire general impressions of their students on the first rating (i.e., their judgment), and this can affect the second rating.

Construct Validity

Construct validity of the BES-2 was examined by (a) its between-ratee discriminability; (b) correlations between the scores on the subscales, and between the scores on each subscale and the total scale; and (c) correlation of the items with their assigned subscale, and with the total scale.

Results of a study with 108 BD students and 102 normally achieving students showed that ratings of the BD students on the BES-2 were significantly different from those of the normally achieving students on all the five subscales ($p < .0001$).

Correlations among the subscales were significantly high, ranging from .57 to .90 ($p < .001$). However, none of them exceeded .90, which is undesirable, because if a correlation between any two subscales is extremely high, they can be identical. Correlation between each subscale and the total scale was also statistically significant at the .001 level.

To determine the extent to which each item contributes to its assigned subscale and to the total scale, scores on the items were correlated with those of the subscales and of the total scale. Sixty-nine of the 76 items significantly correlated with their assigned subscales and with the total scale.

The Portland Problem Behavior Checklist-Revised

Scale Features and Scoring

The revised Portland Problem Behavior Checklist (PPBC) (Waksman, 1983) was designed to provide specific information about children's behavior problems to school and mental health personnel, who should further evaluate the behaviors and determine appropriate services for the children. The battery consists of 29 items, each with five anchor points ranging from 0 (no problem) to 5 (severe problems). A classroom teacher or other teachers who have daily contact with the target children should rate all items.

The PPBC has different forms for four different types of students: (a) males in Kindergarten through grade 6, (b) males in grades 7-12, (c) females in Kindergarten through grade 6, and (d) females in grades 7-12. All of the forms have the same 29 items. However, they differ in item classification: Different number of items belong to different number of subscales across the forms. Items that belong to none of the subscales are classified as "Other Problems," on which the scores are not included in the total. The teacher should rate these items, but the purpose of this category is not specified.

The scoring procedure is simple. A total score, which is needed for evaluation of changes in behavior over time, can be computed by summing all the values except for those on the "Other Problems" category. To identify a child's status, the total score of each subscale is calculated and compared to the normative sample. With each subscale score, the teacher should locate the

child's percentile in the profile chart. A percentile indicates a raw score associated with a specified percentage of scores that are equal to, or lower than the raw score (Levine, 1981).

Norms

The normative sample for the original version of the PPBC consisted of 217 students randomly selected from three elementary schools and one middle school in Portland, Oregon. The sample included 108 males and 109 females, and their grade levels ranged from Kindergarten to grade 8. All students were from regular classes at the time of the assessment, which was conducted during the 1977-1978 school year. Their regular teachers rated the children on the PPBC.

The revised version of the PPBC was normed during the 1982-1983 school year on 306 students (Kindergarten to grade 12) who were randomly selected from 10 schools in Portland, Oregon. This group included 160 boys and 146 girls. Among them, 79 boys and 65 girls were in grades 7-12; 81 boys and 81 girls were in grades K-6. All students were enrolled in regular classes, and had spent a minimum of two months with the teachers who rated their behavior.

Reliability

Waksman (1983) regards split-half reliability as a form of internal consistency. However, he has not specified the procedure for splitting the items, but just reported that the reliability coefficient for all 306 students in the 1982-83 normative sample was .94, "indicating extremely strong internal reliability" (p. 3). The total number of items (i.e., 29) is an odd number, which cannot be divided into two parts equally. Furthermore, all the items are assigned to several subscales, which vary across the four different forms, measuring different factors of problem behavior. Therefore, the split-half reliability should not be used for subscales or total scales.

Test-retest reliability of the PPBC was examined with a group of 239 students randomly selected from Kindergarten to grade 12. The procedures are not described; we are only told that a 2-month interval separated the test from the retest. The reliability coefficient for the entire sample was .81, that for the 117 male students, .85, and for the 122 female students, .78.

Inter-rater reliability was estimated by asking two different teachers to rate the same student on the revised PPBC, and by correlating the two sets of scores. Thirty-seven high school students were tested by 35 teachers. The correlation coefficient for the entire sample was .54, with a correlation of .54 for the male students (n = 20) and a correlation of .49 for female students (n = 17). No information on the statistical significance of the coefficients is available.

Validity

Content Validity

The extent to which the PPBC measures those problem behaviors that require special services is ensured by the item development procedure. Teachers who had referred students to the Multnomah County School Mental Health Program in Portland, Oregon, were asked to identify and state the three most serious behavioral problems found among the students they had referred. They presented a pool of 275 specific problems, and 29 items were generated from them to be included in the final instrument. However, the procedure and criteria for deriving the final items from the pool are not described in the manual.

Construct Validity

To examine the extent of correlation with other measures that are related to the same construct (i.e., student behavior problems), scores on the PPBC were compared to those on other published rating instruments. The correlation between the AML Checklist (Cowen, et al., 1973) and the PPBC scores, obtained on 54 students in grades K-8, was .57. Teacher ratings obtained on 25 students (grades 3 through 8) referred to the School Mental Health Program for school adjustment problems, showed a correlation of .66 between the PPBC and the WPBIC. The PPBC ratings were moderately correlated (r=.49) to those on the Piers-Harris Children's Self-concept Scale (Piers & Harris, 1969), which is designed to measure children's self-descriptions. The scores of 304 students on the PPBC and the Waksman Social Skills Rating Scale (Waksman, 1983), purportedly a measure of students' social skills, showed a correlation of .74 between the two instruments.

Section Summary

By reviewing selected published rating instruments, we exemplified considerations that are usually needed when selecting an instrument for a specific purpose. Those considerations, however, have not been exhaustive.

For example, we have not examined administering procedures because they are fairly straightforward, but some instruments may specify those procedures that play an important role in the rating. Another important consideration that we should make is the language appropriateness of the items. As mentioned earlier, operational or clear statements are critical to eliminate the possibility of rating errors. Statements also should be aligned to the cognitive level of respondents.

Conclusion

Using behavior rating scales can provide useful information that can be incorporated in screening or

grouping individuals, and in designing and validating programs for those who have deviant behaviors. In many cases, published rating instruments can be used or adapted. Care should be taken to select the most appropriate material that will maximally serve the purpose of the rating. However, new instruments must be developed if (a) the individuals to be rated significantly differ from the normative group in terms of their age, ecological or cultural situations, degree of disability, and so forth, and (b) no published instrument is found that contains the target behavioral constructs or domains. The procedures and methods specified in this monograph can be adapted to construct valid and reliable instruments that reduce rating errors. Factor-analytic methods for multi-dimensional scales have been excluded from this discussion. Interested readers may refer to Nunnally (1978) or Gorsuch (1983).

References

Aiken, L. R. (1985). Evaluating ratings on bidirectional scales. *Educational and Psychological Measurement, 45*(2), 195-202.

Aiken, L. R. (1987). Formulas for equating ratings on different scales. *Educational and Psychological Measurement, 47*(1), 51- 54.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

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

Baker, S. B., & Roberts, D. M. The factor structure of the problem-solving inventory: Measuring perceptions of personal problem solving. *Measurement and Evaluation in Counseling and Development, 21*(4), 157-164.

Bens, B. C., & Porter J. W. (1989). A ratio scale measurement of conformity. *Educational and Psychological Measurement, 49*(1), 75-80.

Benson, J., & Rentsch, J. (1988). Testing the dimensionality of the Piers-Harris Children's Self-Concept Scale. *Educational and Psychological Measurement, 48*(3), 615-626.

Bernadin, H. J. (1977). Behavioral expectation scales versus summated scales: A fairer comparison. *Journal of Applied Psychology, 62*(4), 422-427.

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

Bernadin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored rating scales (BARS). *Journal of Applied Psychology, 66*(4), 458-463.

Bird, C. (1940). *Social psychology*. New York: Appleton-Century-Crofts, Inc.

Blanz, F., & Ghiselli, E. E. (1972). The mixed standard scale. *Personnel Psychology, 25*, 185-199.

Boersma, F. J., Chapman, J. W. & Maguire, T. O. (1979). The student perception of ability scale: An instrument for measuring academic self-concept in elementary school children. *Educational and Psychological Measurement, 39*(4), 1035-1041.

Brown, L., & Hammill, D. D. (1978). *Behavior Rating Profile*. Austin, TX: Pro-Ed.

Brown, L., & Hammill, D. D. (1990). *Behavior Rating Profile-Second Edition* (Examiner's Manual). Austin, TX: Pro-Ed.

Brazil, N. & Pollock, Sr., J. H. (1981). Factorial validity of referral preference rating scale. *Educational and Psychological Measurement, 41*(4), 1265-1270.

Burks, H. F. (1970). *School Attitude Survey*. Huntington Beach, California: The Arden Press.

Burks, H. F. (1977). *The Burks' Behavior Rating Scales* (Manual). Los Angeles, CA: Western Psychological Service.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Cassel, R. N. (1962). *The Child Behavior Rating Scale* (Manual). Beverly Hills, CA: Western Psychological Services.

Champion, C. H., Green, S. B., & Sauser, Jr., W. I. (1988). Development and evaluation of short-cut derived behaviorally anchored rating scales. *Educational and Psychological Measurement, 48*(1), 29-41.

Cooley, E., & Ayres, R. (1988). Cluster scores for the Piers-Haris Children's Self-Concept scale: Reliability and independence. *Educational and Psychological Measurement, 48*(4), 1019-1024.

Cowen, E. L., Dorr, D., Clarfield, S., Kreling, B., McWilliams, S. A., Pokracki, F., Pratt, D. M., Terrell, D., & Wilson, A. (1973). The AML: A quick-screening device for early identification of school maladptation. *American Journal of Community Psychology, 1*, 12-35.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, & Winston.

Cronbach, L. J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.

Cronbach, L. J., & Meehl, E. F. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*: 281-302.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*(4), 481-489.

Dickinson, T. L., & Zellinger, P. M. (1980). A compari-

son of the behaviorally anchored rating and mixed standard scale formats. *Journal of Applied Psychology, 65*(2), 147-154.

Dixon, P. N., Bobo, M., & Stevick, R. A. (1984). Response differences and preferences for all-category-defined and end-defined Likert formats. *Educational and Psychological Measurement, 44*(1), 61-66.

Doll, E. A. (1965). *Vineland Social Maturity Scale.* Circle Pines, Minnesota: American Guidance Service.

Edelbrock, C. (1988). Informant reports. In E. S. Shapiro, & T. R. Kratochwill (Eds.), *Behavioral assessment in schools* (pp. 351-383). New York: The Guilford Press.

Edwards, A. I. (1957). *Techniques of attitude scale construction.* New York: Appleton-Century-Crofts, Inc.

Elliott, S. N., Sheridan, S. M., & Gresham, F. M. (1989). Assessing and treating social skills deficits: A case study for the scientist-practitioner. *Journal of School Psychology, 27*(2), 197-222.

Epstein, M. H., & Gadow, K. D. (1986). Teacher ratings of hyperactivity in learning-disabled, emotionally disturbed, and mentally retarded children. *Journal of Special Education, 20*(2), 219-229.

Federal Register. (1977, August 23). [42(163), 42478]. Washington, DC: U. S. Office of Education.

Finley, D. M., Osburn, H. G., Dubin, J. A., & Jeanneret, P. R. (1977). Behaviorally based rating scales: Effects of specific anchors and disguised scale continua. *Personnel Psychology, 30*, 569-669.

Fish, L. J. (1988). Why multivariate methods are usually vital. *Measurement and Evaluation in Counselling and Development, 21*, 130-137.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*(4), 327-358.

Frisbie, D. A., & Brandenburg, D. C. (1979). Equivalence of questionnaire items with varying response formats. *Journal of Educational Measurement, 16*(1), 43-48.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: W. H. Freeman and Company.

Glutting, J. J., Oakland, T., & McDermott, P. A. (1989). Observing child behavior during testing: constructs, validity, and situational generality. *Journal of School Psychology, 27*, 155- 164.

Gorsuch, R. L. (1983). *Factor Analysis.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, S. B., Sauser, Jr., W. I., Fagg, J. N., & Champion, C. H. (1981). Shortcut methods for deriving behaviorally anchored rating scales. *Educational and Psychological Measurement, 41*, 761-775.

Grossman, H. J. (1983). *Classification in mental retardation.* Washington, DC: American Association on Mental Deficiency.

Guilford, J. P. (1954). *Psychometric Methods.* New York: McGraw-Hill.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139-150.

Guttman, L. (1947). The Cornell technique for scale and intensity analysis. *Educational and Psychological Measurement, 7*:247-249.

Hightower, A. D., Work, W. C., Cowen, E. L., Lotyczewski, B. S., Spinell, A. P., & Guare, J. C. (1986). The teacher-child rating scale: A brief objective measure of elementary children's school problem behaviors and competencies. *School Psychology Review, 15*(3), 393-409.

Hightower, A. D., Cowen, E. L., Spinell, A. P., Lotyczewski, B. S., Guare, J. C., Rohrbeck, C. A., & Brown, L. P. (1987). The child rating scale: The development of a socioemotional self-rating scale for elementary school children. *School Psychology Review, 16*(2), 239-255.

Hoffman, R. G., Davis, G. L., & Nelson, K. S. (1988). Factor analysis of Tennessee Self-Concept Scale in an adolescent sample. *Educational and Psychological Measurement, 48*(2), 407- 417.

Jacobs, R., Kafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. *Personnel Psychology, 33*, 595-640.

Kinicki, A. J., & Bannister, B. D. (1988). A test of the measurement assumptions underlying behaviorally anchored rating scales. *Educational and Psychological Measurement, 48*(1), 17-27.

Kinicki, A. J., Bannister, B. D., Hom, P. W., & Denisi, A. S. (1985). Behaviorally anchored rating scales vs. summated rating scales: Psychometric properties and susceptibility to rating bias. *Educational and Psychological Measurement, 45*(3), 535-549.

Kratochwill, T. R., & Shapiro, E. S. (1988). Introduction: Conceptual foundations of behavioral assessment in schools. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools* (pp. 1-13). New York: Guilford Press.

Lam, T. C. M., & Klockars, A. J. (1982). Anchor point effects on the equivalence of questionnare items. *Journal of Educational Measurement, 19*(4), 317-322.

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

Latham, G. P., Fay, C. H., & Saari, L. M. (1979). The development of behavioral observational scales for appraising the performance of foremen. *Personnel Psychology, 32*, 299-311.

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. *Personnel Psychology, 30*, 255-268.

Levine, G. (1981). *Introductory statistics for psychology: The logic and the methods.* New York: Academic

Press.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, No. 140.

MacDonald, L. (1988). Improving the reliability of a maladaptive behavior scale. *American Journal on Mental Retardation, 92*(4), 381-384.

MacDonald, L., & Barton, L. E. (1986). Measuring severity of behavior: A revision of part II of the adaptive behavior scale. *American Journal of Mental Deficiency, 90*(4), 418-424.

McCarney, S., Leigh, J., & Cornbleet, J. (1983). *Behavior Evaluation Scale.* Columbia, MO: Educational Services.

McCarney, S. B., & Leigh, J. E. (1990). *Behavior Evaluation Scale-2.* Columbia, MO: Educational Services.

McGrew, K., & Bruininks, R. (1989). The factor structure of adaptive behavior. *School Psychology Review, 18*(1), 64-81.

McMahon, R. J. (1984). Behavioral checklists and rating scales. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures* (pp. 80-105). Elmsford, N. Y.: Pergamon Press.

Meyers, C. E., Nihira, K., & Zetlin, A. (1979). The measurement of adaptive behavior. In N. R. Ellis (Ed.), *Handbook for mental deficiency: Psychological theory and research* (2nd ed., pp. 431-481). Hillsdale, NJ: Lawrence Erlbaum Associates.

Neeper, R., & Lahey, B. B. (1984). Identification of two dimensions of cognitive deficits through the factor analysis of teacher ratings. *School Psychology Review, 13*(4), 485-490.

Neeper, R., & Lahey, B. B. (1986). The children's behavior rating scale: A factor analytic developmental study. *School Psychology Review, 15*(2), 277-288.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.) New York: McGraw-Hill.

Piers, E. V., & Harris, D. B. (1969). *The Piers-Harris Childrens Self-Concept Scale.* Nashville, Tennessee: Authors.

Quay, H. C., & Peterson, D. R. (1967). *Manual for the Behavior Problem Checklist.* Champaign, Illinois: Children's Research Center.

Raynolds, W. M. (1979). Development and validation of a scale to measure learning-related classroom behaviors. *Educational and Psychological Measurement, 39*(4), 1011-1018.

Rie, E. D., & Friedman, D. P. (1978). *A survey of behavior rating scales for children.* Ohio Department of Mental Health and Mental Retardation.

Robbins, S. B., & Patton, M. J. (1986). Procedures for construction of scales for rating counselor outcomes. *Measurement and Evaluation in Counseling and Development, 19*(3), 131-140.

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.

Saal, F. E., & Landy, F. J. (1977). The mixed standard rating scale: An evaluation. *Organizational Behavior and Human Performance, 18*, 19-35.

Saracho, O. N. (1984). Construction and validation of the play rating scale. *Early Child Development and Care, 17*(2-3), 199- 230.

Schriesheim, C. A., & Novelli, Jr., L. (1989). A comparative test of the interval-scale properties of magnitude estimation and Case III Scaling and recommendations for equal-interval frequency response anchors. *Educational and Psychological Measurement, 49*(1), 59-73.

Schwab, D. P., Heneman III, H. G, & DeCotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology, 28*, 549-562.

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*(2), 149-155.

Spreat, S. (1980). The adaptive behavior scale: A study of criterion validity. *American Journal of Mental Deficiency, 85*(1), 61-68.

Stiffman, A. R., Orme, J. G., Evans, D. A., Feldman, R. A., & Keeney, P. A. (1984). A brief measure of children's behavior problems: The behavior rating index for children. *Measurement and Evaluation in Counseling and Development, 17*(2), 83-90.

Thorndike, R. (1982). *Applied psychometrics.* Boston: Houghton Mifflin.

Thurstone, L. L. (1927). A law of comparative judgment. *The Psychological Review, 34*, 273-286.

Thurstone, L. L. (1928). Attitudes can be measured. *The American Journal of Sociology, 28*(4), 529-554.

Thurstone, L. L. (1946). Comment. *The American Journal of Sociology, 52*, 39-40.

Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude.* Chicago: University of Chicago Press.

Tindal, G. A., & Marston, D. B. (1990). *Classroom-based assessment: Evaluating instrumental outcomes.* Columbus, Ohio: Merrill Publishing Company.

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: John Wiley & Sons, Inc.

Tuckman, B. W. (1988). The scaling of mood. *Educational and Psychological Measurement, 48*(2), 419-427.

Yu, D., Martin, G., Hardy, L., Leader, Clarice, & Quinn, G.(1985). Developing a behavioral assessment system for the mentally handicapped: A behavioral approach. *Canadian Journal of Exceptional Children, 1*(4), 117-123.

Waksman, S. (1983). *The Waksman Social Skills Rating Scale Test and Manual.* Portland, OR: Enrichment Press.

Walker, H. M. (1983). *Walker Problem Behavior Identification Checklist* (Manual). Los Angeles, CA: Western Psychological Services.

Wilson, M. J., & Bullock, L. M. (1989). Psychometric characteristics of behavior rating scales: Definitions, problems, and solutions. *Behavioral Disorders, 14*(3), 186-200.

Wilson, M. J., Moore, A. D., & Bullock, L. M. (1987). Factoral invariance of the behavioral dimensions rating scale. *Measurement and Evaluation in Counseling and Development, 20*(1), 11-17.

Witt, J. C., Cavell, T. A., Heffer, R. W., Carey, M. P., & Martens, B. K. (1988). Child self-report: Interviewing techniques and rating scales. In E. S. Shapiro, & T. R. Kratochwill (Eds.), *Behavioral assessment in schools* (pp. 384- 454). New York: The Guilford Press.

Zbaracki, J. U., Clark, S. G., & Wolins, L. (1985). Children's interests inventory, grades 4-6. *Educational and Psychological Measurement, 45*(3), 517-521.

Zedeck, S., Imparato, N., Krausz, M., & Oleno, T. (1974). Development of behaviorally anchored rating scales as a function of organizational level. *Journal of Applied Psychology, 59*(2), 249-252.