Journal of Counseling Psychology
1987, Vol. 34, No. 4, 481–489

0022-0167/87/$00.75

Scale Construction

René V. Dawis
University of Minnesota

The design, development, and evaluation of scales for use in counseling psychology research are discussed. Methods of scale construction described include the Thurstone, Q-sort, rank-order methods, Likert, semantic differential, Guttman, Rasch, and external criterion methods. Strengths and weaknesses, advantages, and disadvantages are considered, and ways of evaluating newly developed scales are presented. Other issues such as measurement versus statistics, bandwidth versus fidelity, empirical versus rational methods, response bias, and multimethod measurement are discussed.

Scales are ubiquitous features of counseling psychology research. For instance, examination of a randomly chosen issue of the *Journal of Counseling Psychology* (1984, Vol. 31, No. 3) showed that all 12 major articles in the issue reported on studies that involved the use of scales.

As the term is used in counseling psychology research, a *scale* is a collection of items, the responses to which are scored and combined to yield a scale score. Scale scores can be categorized according to level of measurement. At the lowest, nominal level of measurement, scale scores are used to name or designate (identify) the classification categories to which the objects of measurement are grouped. At the ordinal level, scale scores rank order the measured objects along the classificatory dimension. At the interval level, scale scores reflect the relative distances between and among measured objects. At the ratio level, scale scores indicate the absolute distance of any measured object from a true-zero point on the scale. Few, if any, psychological scales are even-interval scales (Thomas, 1982).

Scales can also be classified according to the source of scale score variation, following Torgerson (1958), as stimulus-centered, subject-centered, or response scales. Scale scores in stimulus-centered scales (also called judgment scales) reflect stimulus (item) differences along the measurement dimension. An example would be a life events scale, in which a respondent rates or ranks particular life events in terms of how stressful they are to the respondent. In contrast, for subject-centered scales (also called individual differences scales), scale scores reflect differences among the subjects (respondents) in terms of their standing along the scale's dimension. Personality trait scales of the inventory or questionnaire variety are common examples of subject-centered scales. Lastly, response scales are those for which scale score variation is attributed to both stimuli (items) and subjects (respondents). Scales constructed according to the Rasch scaling methodology (Wright & Masters,. 1982) are examples of response scales.

For the purposes of this article, the term scale will be limited to those instruments that are constructed by researchers in order to obtain quantitative data on variables for which appropriate standardized instruments are not available. Examples of such variables are counselee and counselor perceptions (cognitions), evaluations, feelings, attitudes, plans, and actions (behaviors) as these occur before, during, and after the counseling process. To instrument such variables, researchers have often had to construct their own scales. Typically, such scales rely on the research participant's verbal report, and response by the participant is structured, that is, limited to given choices. This article focuses, therefore, on the construction of verbal, structured scales of the rating, questionnaire, or inventory type. I do not discuss the construction of tests or what Cronbach (1984) calls measures of maximum performance (i.e., ability, aptitude, achievement, knowledge, or skill tests), for which a large literature is available.

The scale construction process may be divided into three stages: design, development, and evaluation. Each stage is discussed in turn.

Scale Design

Designing a scale requires, first of all, some theory of the scale that includes a well-articulated definition of the psychological variable to be measured and indications of how it is to be measured. Definition of the variable depends on the larger theory that impels the research. Definition includes distinctions (what the variable is and what it is not), dependencies (how the variable is a function of more basic or previously defined terms), and relations (how the variable is related to other variables). How the variable is to be measured depends on a number of considerations, such as how best to represent the variable, who the respondents will be, the context and conditions under which the measure is to be administered, and the research design of the study, especially the analyses planned. In short, the theory of the scale should give the scale constructor directions as to scale content (the writing of items) as well as scale format (the type of scale to construct).

Scale Content

A useful preliminary to item writing is to conduct open-ended interviews with representative subjects from the target respondent population. Skillful interviewing can elicit a wide

Correspondence concerning this article should be sent to: René V. Dawis, Department of Psychology, University of Minnesota, 75 East River Road, Minneapolis, Minnesota 55455.

range of statements about the variable in question. The interviewee's own words can then be used in writing the items. Such use can provide a degree of authenticity that in turn can contribute to the scale's validity. For example, a scale to be filled out by clients to describe the counselor's behavior would be much more acceptable and credible to the clients if it were expressed in their (lay) language rather than in the more technical (if more precise) language of theory. Use of respondents' own words will also mean that readability of the scale will be less of a problem. Whether respondents' own words are used or not, it is good practice to check on the readability level of the scale to make sure that it is appropriate to the level of the respondent population. Useful hints on writing readable items are given by Payne (1951) and Fiske (1981).

The initial pool of items can be written to be homogeneous or heterogeneous in content. The scale design should indicate what degree of content heterogeneity is desired, based on the theory of the scale. A useful aid in this regard is to look at the scale data matrix as a two-factor, completely crossed with no replication analysis of variance design, in which the two factors are items and respondents. One can then see that, depending on the researcher's purposes, the scale can be so constructed as to maximize item effect only, respondent effect only, or item-by-respondent interaction. Maximizing item effect will require heterogeneous items; minimizing it will require homogeneous item content.

If items are explicitly derived from hypotheses from the larger theory, it might be useful to do a "back translation" (Smith & Kendall, 1963). That is, competent judges who were not involved in the writing of the items could be asked to assign the items back to the hypotheses or hypothesized categories. Back translation can be a useful check on the coverage of the content domain as outlined by the design of the scale.

Scale Format

Items in structured verbal scales typically consist of a stimulus part (the item stem) and a response part (the response choices). Item stems may consist of full sentences, phrases, or even single words. They may describe some attribute of an object (e.g., "The counselor appears trustworthy"), or the state of the object ("The counselor is passive"), or some event involving the object ("The counselor is reflecting the client's feelings"), to varying degrees of specificity or generality. Item stems ordinarily consist of single components, but may have two or more components (as in paired comparison or multiple rank-order scales).

Response choices in structured verbal scales vary in their underlying measurement dimension (e.g., agree–disagree, like–dislike, important–unimportant). They also vary in response format. Rating response formats differ in the number of scale points (choices) given the respondents (2-, 3-, or 5-point scales are the most common), and in the way scale points are anchored. Anchors can be words (*yes–no, true–false*), phrases (*strongly agree–strongly disagree*), or more extended prose as in behaviorally anchored scales (e.g., Campbell, Dunnette, Arvey, & Hellervik, 1973). Rating scales may be anchored at each scale point or only at selected scale points (e.g., at the ends and the middle of the scale). Response choices may be unweighted (scored with 0, 1 weights), or weighted using multiple weights. Rating response formats may be one-sided (zero to positive or to negative values) or two-sided (with both positive and negative sides of the continuum).

Ranking response formats are fewer in number, differing only in the number of elements ranked within an item (e.g., paired comparison, multiple rank orders such as triads and tetrads, or, at the extreme, a single ranking of all elements). Ranking response formats use ranks rather than weights as scores and by convention, ranks are ordered in a manner opposite that of weights in the rating response format, that is, the lower the number, the higher the rank.

In choosing a scale format, the general rule might be to choose the simpler format. However, there are other considerations: More complex formats might make the task of filling out the scale more interesting for the more experienced or knowledgeable respondent. When rating response formats are used, more scale points are better than fewer, because once the data are in, one can always combine scale points to reduce their number, but one cannot increase that number after the fact. Also, more scale points can generate more variability in response, a desirable scale characteristic if the response is reliable. Inordinate use of the middlemost scale point can be avoided by eliminating that scale point, that is, by using an even number of scale points. This has the further advantage of ensuring that the underlying dimension will be linear or can be made linear. At times rank ordering may be easier to do than rating, but use of ranking response formats may place limits on the statistical analysis of the data. Finally, the amount of space available for the scale (e.g., in an extended questionnaire) might preclude the use of certain formats.

Scale Development

Scale development consists of collecting data with the use of a preliminary form and analyzing the data in order to select items for a more final form. ("More final" is intended to indicate that the process might have to undergo one or more iterations depending on the results of the evaluation stage.) It is always useful to conduct a small N pilot study before the main data collection effort. The pilot study can be used to check out such nuts-and-bolts points as how easily the scale instructions are followed, how well the scale format functions, how long the scale takes to complete, and especially, how appropriate the scale items are for the target respondent population.

As a rule, the development sample should be representative of the target respondent population. There can be exceptions, however; for example, in developing stimulus-centered scales, one could use a sample that is more homogeneous than samples from the target population.

At the heart of scale construction is the scaling method used to select items. Several methods are described, grouped according to the type of scale (stimulus-centered, subject-centered, or response) with which they are typically associated. A fourth group of methods, the external criterion methods, which select items on a different basis, are also described.

Stimulus-Centered Scale Methods

Because counseling psychology is concerned with the individual client, one might expect more frequent use of stimulus-centered scales than apparently is the case. How a particular client scales stimuli (e.g., stressfulness of life events, preference for occupations) regardless of how others do it should be just as significant for counseling as how the individual compares with others, if not more so. Stimulus-centered scales would appear to be particularly appropriate to use in monitoring the progress of the client during counseling.

The prototypic method for developing stimulus-centered scales was the Thurstone method (Thurstone & Chave, 1929). From this method developed the popular Q sort. Rank-order methods are also frequently used to construct stimulus-centered scales. Brief descriptions of these methods follow:

The Thurstone method. Thurstone's groundbreaking insight was that questionnaires could be constructed as scales by the application of the methods of psychophysics. The Thurstone method proceeds as follows:

1. A large number of statements (say, 200 to 300) are written about the construct, to represent the range of the construct.

2. A number of judges (say, 20 to 30) are asked to sort the items with respect to the underlying measurement dimension and to assign an appropriate scale value (scale point on the numerical scale) to each item. An 11-point scale is typically used.

3. The central tendency and variability of scale values assigned to it are computed for each item.

4. On the basis of their average scale values, two or three items with the lowest variabilities are selected to represent each scale point. Thurstone scales typically have 22 items.

After the items have been selected, they are arranged in random fashion in a questionnaire. Respondents are instructed, for instance, in the case of an attitude scale, to identify those items that they endorse. (Similar instructions can be given for other types of scales, for example, identifying the items descriptive of self, or of another person being rated, or of the events being observed.) The scale score is calculated as the average of the scale values of the endorsed items.

The Thurstone method, although a historic methodological breakthrough, has not found much favor with scale constructors, and is practically unheard of in counseling psychology. Much better known is its derivative, the Q sort.

The Q-sort method. The Q-sort method (Stephenson, 1953) has been used extensively in personality research, especially in research on the self-concept. The Q-sort method starts with a fixed set of stimuli (e.g., self-descriptive statements). The respondent is asked to sort the stimuli along a scale according to scale value (e.g., least to most descriptive). To ensure variability in the scores and to forestall response biases such as central tendency or leniency, the respondent might be asked to force the stimuli into a distribution, for example, for a 5-point scale, a 7%-24%-38%-24%-7% distribution.

The Q sort is useful in situations in which multiple response roles (positions) are taken with respect to the same set of stimuli (e.g., in self-concept research, "How I actually am," "How I would like to be," "How others see me," etc., are response roles that can be used with the same set of self-descriptors). Q-sort data are typically used in Q correlation (correlation between persons across variables) or in O correlation (correlation between occasions across variables). They may also be used in ordinary *R* correlation, unless the forced distribution method is used. In the latter case, the Q-sort scores will be ipsative. Ipsative scores (Clemans, 1966) are those in which the scores for an individual are distributed around that individual's mean and not around the population mean. Ipsative scores are not on a common scale for all individuals and therefore cannot be used in analyses that assume a common scale, for example, correlating variables across individuals, factor analysis, or analysis of variance. However, they would be appropriate in correlating individuals across variables (i.e., in Q correlation).

Rank-order methods. The two frequently used rank-order methods are the paired comparison method and the ranking method.

In the paired comparison method (Guilford, 1954), each stimulus (e.g., person, object, event, state, or condition) is paired with every other stimulus. The respondent's task is to select one stimulus from each pair on the basis of the scaling dimension, that is, the basis of comparison. From the number of times each stimulus is chosen, the stimuli can be rank ordered with more precise information than if all of the stimuli were just rank ordered in the usual way. (The additional information comes from circular triads, i.e., where A is chosen over B, B over C, and C over A. Such information is not obtained in ordinary ranking.)

Each stimulus' "score" (number of times chosen) can also be converted to a *z* score, using the normal curve table. Such *z* scores would be ipsative. The ipsative character of these *z* scores can be minimized by calibrating each individual's scores to that individual's zero point. This zero point can be ascertained (for each individual) by adding an absolute judgment scale (a two-categoried scale; see Guilford, 1954, pp. 171–173).

Because the number of pairs increases rapidly with increase in number of stimuli (for *n* stimuli, the total number of pairs equals $n[n - 1]/2$), the paired comparison method becomes impractical when more than 20 stimuli are involved. For such situations, the method of multiple rank orders (Gulliksen & Tucker, 1961) can be used, in which, instead of presenting stimuli in pairs, they are presented in blocks of threes (triads) or more, but in such a manner that each stimulus is paired only once with every other stimulus. (Special designs are necessary to accomplish this. See Guilliksen & Tucker, 1961.) If collected in this way, the data from the multiple rank orders can be reduced to paired comparisons, and then scaled as paired comparisons.

At the other extreme to paired comparisons is the ranking method. Ranking can be used with any number of stimuli. For small numbers, the instructions are straightforward. For large numbers of stimuli (i.e., more than 20), the reliability of the ranking can be improved by using the alternation ranking procedure, in which the respondent alternates between picking the highest and lowest ranks (i.e., first, the first ranked; next, the last ranked; then, the second ranked; then, the next to the last ranked; the third ranked, etc.). As with paired comparison data, ranking data can also be converted to z scores (Guilford, 1954).

Ranking data, that is, rank scores, whether obtained by the paired comparison, multiple rank orders, or ranking method, should be analyzed by using nonparametric statistics (Siegel, 1956), especially rank-order statistics. When converted to z scores with a zero point, however, the data can be analyzed with the use of ordinary parametric statistics.

Subject-Centered Scale Methods

Subject-centered scales are probably the kind of scale in most frequent use in counseling psychology research. Individual differences in both the clients and the counselors are thought to account for significant portions of counseling outcome variance. Also, possibly because individual differences variables are among the most easily accessible to researchers, much effort has been put into constructing and developing subject-centered scales.

The classic method for developing subject-centered scales is the Likert method. Refinements in the method have been introduced via factor analysis. A variant of the method, the semantic differential, has proven quite useful. These methods are described below.

The Likert method. Just as Thurstone saw the application of psychophysical methods to scaling nonsensory stimuli, so did Likert (1932) see the application of psychometric methods to scaling nonability-test items. The Likert procedure can be described as follows:

1. A number of items are written to represent the content domain. Five-point anchored rating scales are typically used as response choices for each item (hence, the mistaken use of *Likert* to refer to the 5-point-rating item format). Scoring weights from 1 to 5 are assigned to the five rating-scale points. Direction of scoring (whether 1 or 5 is high) is immaterial provided it is consistent for all items.

2. The items are administered to a large group of respondents (*N* of at least 100). Each respondent's item rating choices are scored and the item scores summed to constitute the respondent's total score.

3. Items are selected according to their ability to discriminate between high and low scorers on total score. Likert used a group-difference procedure (difference in item means between high-scoring and low-scoring groups, e.g., uppermost 25% and lowermost 25%). One could also use an item–total-score correlation procedure, as is currently done in ability test construction. Maximizing item–total-score correlation will also maximize the scale's internal consistency reliability coefficient (coefficient alpha). Computer programs (e.g., the Statistical Package for the Social Sciences Reliability program) are available for use in this connection.

4. The best discriminating items are then selected to constitute the scale, and the scale score is obtained by summing the item scores for the selected items. At this point, scale scores can be treated as normative scores (i.e., transformed to standardized scores, used to determine percentile equivalents for specific populations, etc.).

Of all the scale construction methods, the most convenient for researchers is the Likert method because it can be employed with the use of ordinary SPSS programs. To implement the Likert method requires only (a) computing total score, (b) computing item–total-score correlations, and (c) computing alpha reliability for the final set of items. Incidentally, reliability should be computed for every research use of Likert scales, not just at scale development, because reliability is a function not only of the scale but also of the respondent sample.

Unfortunately, not all scales that are purported to be Likert-type scales are constructed according to the Likert procedure. They only look like Likert scales because of the use of the 5-point rating response format (Triandis, 1971). If, in such scales, the correlation of the items with total scale is not high, then the interpretation of the scale score is problematic.

Use of factor analysis. Factor analysis is a data reduction technique in which a large set of variables is reduced to a smaller set without much loss of information. The technique can be used to select items for Likert-type scales in the following way:

1. The items in the item pool are intercorrelated.

2. The item intercorrelation matrix is subjected to a principal components analysis. (This requires the use of the principal axis solution, with unities in the diagonal, and extracting only the first factor.)

3. The items with the highest loadings are selected for the scale. *Highest loading* can be defined in an absolute sense (e.g., at least .707 or .50, which would represent 50% and 25%, respectively, of the item variance) or in a relative sense (the loading squared, as a proportion of the communality, e.g., no less than 50%).

4. There may be instances in which certain items are essential to the definition of the scale but are not found among the highest loading ones, that is, are not selected by this procedure. In this case, the scale constructor can go back to the original item intercorrelation matrix and eliminate all items that correlate below a given level (e.g., .30) with the essential defining items. The reduced matrix can then be factor analyzed.

5. When a content domain represented in an item pool is thought or assumed to be multidimensional, factor analysis can be used to construct several scales at the same time. The procedure is the same as above, except that more than one factor (component) is extracted. An additional step, factor rotation, is usually required to find a best (simple structure) solution, the procedure most frequently preferred being orthogonal rotation to a varimax criterion. A scale is then constructed for each factor, with items selected as described above. If an item is selected for more than one scale, the researcher can choose (a) to assign it to the scale with the highest loading, (b) to assign it to all the scales for which it was selected, or (c) to leave it out altogether. Choices (a) and (c) waste some information, but choice (b) will contribute to an artifactual interscale correlation that is undesirable.

As with Likert scales, all scales developed via factor analysis should be evaluated for reliability each time they are used.

The semantic differential. The semantic differential (Osgood, Suci, & Tannenbaum, 1957), like the Likert, makes use of the rating response format. Unlike the Likert, which uses only one rating dimension for all items in a scale, the semantic differential uses several rating dimensions for rating the same item or stimulus object. Semantic differential rating dimensions are typically bipolar, anchored at both ends with contrasting adjectives, with a 7-point rating continuum. Provided that response distributions are not forced, semantic differential data can be treated like any other rating data.

Response Scale Methods

If stimuli can be assigned scale scores and subjects can be assigned scale scores, the next logical development should be to develop scale construction methods that assign scale scores to both subjects and stimuli. Such development has been going on (e.g., Coombs, 1964) but has been the province mainly of psychologists interested in scaling models and psychological modeling. Only relatively recently has response scale development had an impact on instrument construction in applied psychology (e.g., Lord & Novick, 1968). It has had practically no impact on counseling psychology research.

For the sake of completeness, however, and to illustrate the response scale approach, one of the earliest and more influential scaling methods—Guttman's scalogram technique—will be described. A more recently developed technique, the Rasch (1960) method, will also be briefly discussed.

The Guttman method. Guttman's (1944) concern was the property of unidimensionality in a scale. With a unidimensional scale, according to Guttman, knowledge of the respondent's scale score should permit the reproduction of the respondent's item score pattern. In a unidimensional scale, the items can be arranged in order (of endorsement or descriptiveness or whatever the underlying dimension is) in such a way that positive response to an item (e.g., *agree*, in an attitude scale) should imply positive response to all other items lower down the scale, and conversely, negative response to an item should imply negative response to all other items higher up the scale. To ascertain unidimensionality, Guttman developed the scalogram technique.

Suppose we had a unidimensional attitude scale that was administered to a group of individuals. The scalogram technique would call for the data to be displayed as follows: Items are displayed as columns and ordered (from left to right) according to endorsement level from the most to the least endorsed item. Individuals are displayed as rows and ordered (from top to bottom) according to total score, from highest to lowest score. If the test were perfectly unidimensional, then the scalogram would show an orderly stepwise progression of endorsement for both the individuals and the items. Any exceptions to this expectation can be easily seen in a scalogram display, and the number of exceptions can be expressed as a proportion of the total matrix (N individuals × m items). Guttman (1944) defines a *coefficient of reproducibility* as 1 minus the proportion of exceptions, where 1.00 means that the response pattern for any given scale score can be reproduced perfectly.

When the coefficient of reproducibility is not high (e.g., below .9 or .8), the scalogram display will reveal the items that do not conform to expectation. After removing these items, the coefficient of reproducibility is recalculated, and the process repeated until the desired level of the coefficient is attained. Sometimes it may also be necessary to eliminate some aberrant individuals whose responses do not conform to the expected pattern. (This underscores the fact that response is a function not just of the scale or instrument but also of the respondent population. Aberrant individuals might be hypothesized to belong to a different population insofar as the scale is concerned.)

The classic Bogardus (1928) social distance scale illustrates what Guttman had in mind. Respondents were asked whether they would be willing to admit members of a race or nationality group (a) to close kinship by marriage, (b) to membership in their club, (c) to their streets as neighbors, (d) to employment in their occupation, (e) to citizenship in their country, (f) only as visitors to their country, or (g) whether they would exclude them completely from their country. Admitting individuals at one level implies admitting them at lower levels but does not imply admitting them at higher levels.

The Rasch method. The Rasch model, one of a group of models originating from item response theory, was initially developed in connection with the construction of ability tests. The model expresses Guttman's basic ideas in a probabilistic manner, as follows: (a) Given any item, a person of higher ability should have a higher probability of getting the item right than would a person of lower ability, and (b) given any person, an item of lower difficulty should be solved (gotten right) with a higher probability than would an item of higher difficulty. The model has since been extended to the construction of nonability measures (e.g., attitude scales) by, among others, Wright and Masters (1982).

The Rasch model postulates that item response is a function of two parameters, an item parameter and a person parameter. As examples: For ability tests, the parameters would reflect item difficulty and person ability; for attitude scales, item endorsement, and person attitude; for interest measures, item liking (liking for an item) and person interest. The parameters are estimated from the item-by-score matrix (persons with the same scores are grouped together). Parameters are estimated from the data, given that the model is true (i.e., with the model as the premise). The data's fit to the model can be assessed, and if the fit is poor, one infers that the model's assumptions have not been met.

If the fit is acceptable, the data can be improved by eliminating the items that show a poor fit (and in theory, the persons that have a poor fit, as well). Thus, by eliminating poorly fitting items, the refined scale is assumed to be unidimensional. (The reader will note the similarity to the Guttman technique.)

Calculation of parameter estimates for the Rasch model is typically done via computer, although hand calculation methods are also available (Wright & Masters, 1982).

All of the scale development methods described thus far make use of the item-by-person data matrix in determining which items to retain in, or eliminate from, the scale. A final group of methods makes use of external criteria and the relation of items to external criteria in determining which items to select. These methods were developed in the context of the practitioner's problem of predicting outcomes (e.g., in vocational choice and personnel selection). For these methods, the choice of criterion (or criteria) is all-important because it preordains the items that are selected.

External criterion methods. Item selection, again, is the key question in scale construction by external criterion methods. The three most-used methods of item selection are (a) the group difference method, (b) the item validity method, and (c) the multiple regression method. It is assumed that the criterion variable has been selected and that an adequate measure of it is available. Criterion variables typically reflect whatever it is that psychologist-practitioners are trying to effect, for example, client satisfaction, client choice, client behavior. (To simplify discussion, a single criterion variable is assumed, although a scale can be constructed to predict to multiple criteria.)

In the group-difference method, items for the scale are selected according to the difference in mean item scores between two groups, a high criterion group and a low criterion group, or, alternatively, a criterion group (whose members meet one or more criteria) and a reference group (a baseline, or unselected, or typical-population group). The larger the mean difference, the more definitely the item should be selected for the scale. The size of the difference can be used to give differential weights to items and response choices (Strong, 1943), but when the number of items in the scale is large (20 or so), unit weights (0, 1) do just as well as differential weights (Clark, 1961).

Note that the group difference method is similar to Likert's original method. What differs is that Likert used an internal criterion (total score on the undeveloped scale), whereas the present method uses an external criterion. Otherwise, the statistical procedures are very much the same.

The item validity method is also similar to the Likert method except that instead of the Likert's item–total-score correlation, the external criterion method uses the correlation between item score and criterion score as the basis for item selection.

A more sophisticated external criterion method of item selection involves the use of multiple regression. The criterion variable is regressed on the items, with items being added to the regression equation one at a time, depending on the amount of explained variance the item contributes. This method tends to select items that correlate highly with the criterion and lowly or not at all with one another.

Scales developed by external criterion methods tend to be heterogeneous in content, because most criteria tend to be heterogeneous or multidimensional. If this is so, determining internal consistency reliability may not be appropriate for scales constructed by these methods. Rather, immediate test–retest or alternate-forms methods should be used to ascertain reliability.

Because external criterion methods tend to capitalize on chance (i.e., sample idiosyncracies), three preventive steps should be taken: (a) The contrast groups (high vs. low, criterion vs. reference) should be large (Strong, 1943, used groups of at least 400); (b) the mean item score differences or item–total-score correlations should be of practical, not just statistical, significance; and (c) the developed scale, after item selection, should be cross-validated, that is, tried out on new samples from the same population as the development sample. Cross-validation, to see if the group differences or correlations hold up, is of the utmost importance in scale construction by external criterion methods.

The fact that external criterion methods are designed to maximize the prediction of criteria is both their strength and their weakness. When the purpose of constructing a scale is to predict to a given criterion, an external criterion method is still unsurpassed as the method of choice. However, a scale that is developed to predict to an external criterion is only as good as the criterion at the time of scale development. If the criterion changes with time (e.g., a shift in emphasis in the criterion from quantity to quality), then the scale can become obsolete. If the criterion happens to be biased against one sex or one ethnic group, then the scale will also be so biased. With new criteria, new scales may have to be constructed, although not before the old scales have been tried and found wanting. Otherwise, a seemingly never-ending series of new scale construction may result. For this reason, use of external criterion methods may require prior resolution of the criterion problem on theoretical as well as on practical grounds.

Scale Evaluation

Scales, as measuring instruments, are evaluated primarily on the basis of two criteria: reliability, or the proportion of scale score variance that is not error variance, and validity, or the proportion of scale score variance that accurately represents the construct or the proportion of criterion variance that is predicted by the scale. These two criteria are complex concepts, and a full discussion of them will not be attempted. However, certain points need to be made in connection with the evaluation of newly constructed scales. (A necessary reference for all scale constructors is the American Psychological Association's *Standards for Educational and Psychological Testing*, 1985.)

That different kinds of reliability estimates may be required for different kinds of scales has already been mentioned. For stimulus-centered scales, the reliability concern is whether on immediate retest the stimuli (items) will be rank ordered in the same way by the same person. The variance of the difference scores between test and retest would be indicative of error variance. For subject-centered scales, the concern is whether individuals are rank ordered in the same way on immediate retest. Variability in individuals' standing would be error variance. For trait scales, reliability refers to the stability of scores (or rank-order standing) over considerable lengths of time. This assumes that individuals are mature on the trait (i.e., developmentally in the stage when the trait is presumed to be stable). For state scales, reliability is the ability of the scale accurately to reflect changes in direction or intensity, or both, in the state being measured. For homogeneous scales, internal consistency reliability is appropriate; for heterogeneous scales, it is immediate test–retest or alternate-forms reliability.

Also, because reliability is a function of sample as well as of instrument, it should be evaluated on a sample from the intended target population—an obvious but sometimes overlooked point.

With respect to validity, although the concept continues to evolve (Messick, 1981; Embretson, 1983), certain practices have come to be accepted as mandatory. One of these is the use of the multitrait–multimethod design (Campbell & Fiske, 1959) to evaluate a scale. At the very least, the scale constructor should compare the new scale with the best competing scale and with a measure of a construct that clearly contrasts with the new scale (e.g., a positive affect scale against a negative affect scale).

It is also common—and good—practice to ascertain the correlates of the scale (e.g., age, sex, experience). It is even better if the expectations about correlates are given by theory. In ascertaining such theory-derived correlates, the *nomological net* (Cronbach & Meehl, 1955) that characterizes the construct is given concrete definition. However, such a network of correlations and other relations only delimits the scale's *nomothetic span* (Embretson, 1983). If the scale is purported to be a measure of a construct, validation studies would have to identify the mechanisms that produce the scale scores and relate these mechanisms to the construct (i.e., do what Embretson calls *construct representation*).

The practical validity or utility (usefulness in professional practice) of a scale is still mainly a matter of predicting to criteria, either concurrently or subsequently measured. The number and range of criteria to which a scale can predict delineate its utility. The most useful scales in professional psychological practice (e.g., Minnesota Multiphasic Personality Inventory [MMPI], Strong–Campbell Interest Inventory [SCII]) are characterized by the large number and wide range of criteria for which the scales are valid predictors.

Prediction to a criterion can be evaluated in two ways: by correlation (proportion of criterion variance accounted for) or by hit-rate (proportion of predicted positives who are true positives). The two are related in the Taylor-Russell tables (Taylor & Russell, 1939), which show what the hit rate would be as a function of the validity coefficient, the base rate, and the selection ratio. Hit rate data are much more concrete and much more useful to the counseling psychology practitioner than are correlation data.

Although reliability and validity concerns are of the essence, there are other less important (but nonetheless, important) considerations. Some of these have been mentioned, for example, administrative concerns. Another concern is the character of the score distribution generated by the scale—in part, a function of the respondent sample. Most users would prefer a scale that ordinarily produces a reasonably normally distributed set of scores. However, if the scale were to be used for diagnostic purposes, a user might prefer one that generates a skewed distribution, the direction of skew depending on whether low scores or high scores are diagnostic. Scales, like ability tests, can be so constructed as to produce the shape of score distribution that is desired, by selecting the appropriate items. Another concern is that the scale produce sufficient score variation to be useful, that is, produce unattenuated correlations. An old rule of thumb is that the coefficient of variation (standard deviation divided by the mean) should be between 5% and 15% (Snedecor, 1946, p. 47).

A final concern is a practical one: Is this scale necessary? That is, are there other, less expensive ways of getting the same information or the same measurements? This concern could also be a matter of social sensitivity: Are there other, less intrusive ways of getting the same information or measurements?

Other Issues

A number of other issues continue to be controversial or, at least, matters of concern for scale constructors.

1. Measurement versus statistics. This is an old and continuing debate that has recently been renewed (Gaito, 1980; Townsend & Ashby, 1984). In brief, the proponents of measurement hold that level of measurement (nominal, ordinal, interval, ratio) constrains the kinds of statistical procedures that can be applied to the numerical data. The proponents of statistics maintain that, "(t)he numbers do not know where they come from" (Lord, 1953, p. 751), that the level of measurement is not a constraining factor. Those who accept the latter view tolerate the use of parametric statistics with scores from quasi-interval scales that actually are at the ordinal level of measurement, a common practice that is criticized by proponents of the former view.

2. Bandwidth versus fidelity (Cronbach & Gleser, 1965). This is the scale constructor's dilemma that can be illustrated as follows: Suppose, for whatever reason, you are limited to 30 items. Do you construct a scale that yields a single, highly reliable score from 30 items or a scale that can yield three independent scores from three 10-item subscales, even if these subscale scores are of marginal reliability? The trade-off is reminiscent of an older one called the *attenuation paradox* (Loevinger, 1954), which identified a trade-off between reliability and validity. That is, high reliability is achieved at the expense of validity and high validity is achieved at the expense of reliability. Ways out of the paradox have been suggested (Humphreys, 1956).

3. Empirical versus rational scales. Conventional wisdom in applied psychology used to hold that empirical (external criterion) scales were the more valid, whereas rational (internal criterion, intuitive) scales were the more reliable. The opinion—or at least the part about the superiority of empirical scales with respect to validity—has now been challenged (Ashton & Goldberg, 1973; Goldberg, 1972; Hase & Goldberg, 1967; Hornick, James, & Jones, 1977; Jackson, 1975).

4. The reference group problem. In the use of the external criterion method of scale construction, what should constitute a reference group? The answer to this question may seem obvious (i.e., a proportionately representative sample of the population), but more careful examination will show that the answer is not so obvious. What is the referent population? The general adult population? A particular age group or sex group? On what variables should there be proportionate representation? Is equal representation better? The constitution of the reference group is important because the scoring key (items selected, weights for response choices) can change with the change of the reference group. One solution is to bypass the reference group (as Kuder, 1977, did) and use the criterion group's responses to develop the scoring key.

5. Response bias. Ratings—whether self- or other-descriptive, general (abstract) or specific (behavioral), or other kinds—are susceptible to certain response biases on the part of respondents. A response bias is a response tendency that operates in all rating situations, regardless of the context. At least three types of bias can be identified: (a) level bias, or the tendency to locate the mean of the ratings high on the scale (leniency or generosity), low on the scale (strictness or severity), or in the middle (central tendency); (b) dispersion bias, or the tendency to constrain or to expand the distribution of ratings (use of a small segment of the scale vs. use of the full range); and (c) correlation bias, which applies when several rating scales, dimensions, or items, i.e., variables, are involved. In such a situation, a common tendency called the *halo effect* results in the high correlation of variables. The opposite tendency, resulting in low or zero correlations, is rarely, if ever, observed. Most of the controversy concerns correlation bias, with some (e.g., Jackson & Messick, 1961) arguing for its removal in every case, and others (e.g., Block, 1965) arguing that such correlations are not necessarily bias and could be veridical. In any event, a large first principal component in rating data is a common finding, sometimes contrary to the expectations of the scale constructor.

6. Multimethod measurement. It is conventional wisdom nowadays to advocate the use of more than one method of measuring any construct. Such a recommendation may overlook the possibility that a change of method can change what it is that is being measured. In other words, method of measurement should be an integral part of the definition and explication of a construct.

7. Direction of measurement. Seemingly bipolar variables sometimes pose problems for scale constructors in that scaling in one direction can result in a measure that does not correlate highly with another that is scaled in the opposite direction. Some constructs, such as masculinity–feminity and positive versus negative affectivity, initially construed as bipolar but unidimensional, have now been redefined as bidimensional. Others such as flexibility–rigidity, while still construed as unidimensional, nevertheless require two different scales for measurement at each pole. These phenomena underscore the need for an adequate theory of the construct to start with, but also for theory to be open to modification in the light of data.

8. A final issue. The demand for some quantitative measure of the multitude of process or outcome variables in counseling psychology research, coupled with the convenience of putting together a structured verbal scale, especially one of the Likert type, has led to the almost exclusive or even automatic use of such measures in our field. That researchers are quantifying their variables through the construction and use of such scales is laudable. That such scales have become the instrument of choice in our field is somehow worrisome. Just as we have been criticized for having developed a psychology of the college sophomore, may we not now be accused of having developed a psychology of the Likert scale response?

A concluding note. In scale construction, as in much of human endeavor, there can be no single "best" method. One method may be best for one research problem but not for another. Purpose, context, and limitations on the researcher have to be taken into account. Trade-offs in advantages and disadvantages seem to be the rule. A hybrid approach, tailored to the situation, might be better than any of the standard approaches discussed here. Researchers should not be reluctant to experiment with different scale construction approaches—and should report their results, so that the rest of us can find out what method is best.

References

Ashton, S. G., & Goldberg, L. R. (1973). In response to Jackson's challenge: The comparative validity of personality scales constructed by the external (empirical) strategy and scales developed intuitively by experts, novices, and laymen. *Journal of Research in Personality, 7,* 1–20.

Block, J. (1965). *The challenge of response sets: Unconfounding meaning, acquiescence, and social desirability in the MMPI.* New York: Appleton-Century-Crofts.

Bogardus, E. S. (1928). *Immigration and race attitudes.* Lexington, MA: Heath.

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

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. (1973). The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology, 57,* 15–22.

Clark, K. E. (1961). *The vocational interests of non-professional men.* Minneapolis: University of Minnesota Press.

Clemans, W. V. (1966). An analytical and empirical examination of some properties of ipsative measures. *Psychometric Monographs (14).*

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York: Harper & Row.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions.* Urbana: University of Illinois Press.

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

Embretson (Whitely), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Fiske, D. W. (Ed.). (1981). *Problems with language imprecision.* San Francisco: Jossey-Bass.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87,* 564–567.

Goldberg, L. R. (1972). Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monographs, 72*(2).

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Gulliksen, H., & Tucker, L. R. (1961). A general procedure for obtaining paired comparisons from multiple rank orders. *Psychometrika, 26,* 173–183.

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

Hase, H. D., & Goldberg, L. R. (1967). Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin, 67,* 231–248.

Hornick, C. W., James, L. R., & Jones, A. P. (1977). Empirical item keying versus a rational approach to analyzing a psychological climate questionnaire. *Applied Psychological Measurement, 1,* 489–500.

Humphreys, L. G. (1956). The normal curve and the attenuation paradox in test theory. *Psychological Bulletin, 53,* 472–476.

Jackson, D. N. (1975). The relative validity of scales prepared by naive item writers and those based on empirical methods of personality scale construction. *Educational and Psychological Meas-*

urement, 35, 361–370.

Jackson, D. N., & Messick, S. (1961). Acquiescence and desirability as response determinants on the MMPI. *Educational and Psychological Measurement, 21,* 771–790.

Kuder, F. (1977). *Activity interests and occupational choice.* Chicago: Science Research Associates.

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

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51,* 493–504.

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8,* 750–751.

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin, 89,* 575–588.

Osgood, C. E., Suci, C. J., & Tannenbaum, P. H. (1957). *The measurement of meaning.* Urbana: University of Illinois Press.

Payne, S. L. (1951). *The art of asking questions.* Princeton, NJ: Princeton University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danmarks Paedogogiske Institut. (Chicago: University of Chicago Press, 1980).

Siegel, S. (1956). *Nonparametric statistics for the behavioral science.* New York: McGraw-Hill.

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

Snedecor, G. W. (1946). *Statistical methods.* Ames: Iowa State College Press.

Standards for educational and psychological testing. (1985). Washington, DC: American Psychological Association.

Stephenson, W. (1953). *The study of behavior.* Chicago: University of Chicago Press.

Strong, E. K., Jr. (1943). *Vocational interests of men and women.* Stanford, CA: Stanford University Press.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565–578.

Thomas, H. (1982). IQ, interval scales, and normal distributions. *Psychological Bulletin, 91,* 198–202.

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

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

Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin, 96,* 394–401.

Triandis, H. C. (1971). *Attitude and attitude change.* New York: Wiley.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: Mesa Press.

Received April 7, 1986
Revision received October 6, 1986
Accepted October 14, 1986 ■

Schmitt Appointed Editor of the *Journal of Applied Psychology*, 1989–1994

The Publications and Communications Board of the American Psychological Association announces the appointment of Neal Schmitt, Michigan State University, as editor of the *Journal of Applied Psychology* for a 6-year term beginning in 1989. As of January 1, 1988, manuscripts should be directed to

Neal Schmitt
Department of Psychology
Psychology Research Building
Michigan State University
East Lansing, Michigan 48824

Manuscript submission patterns for the *Journal of Applied Psychology* make the precise date of completion of the 1988 volume uncertain. The current editor, Robert Gulon, will receive and consider manuscripts until December 31, 1987. Should the 1988 volume be completed before that date, manuscripts will be redirected to Schmitt for consideration in the 1989 volume.