CHAPTER 1

Criteria for Scale Selection
and Evaluation

John P. Robinson, Phillip R. Shaver, and Lawrence S. Wrightsman

The original idea for this handbook of attitude and personality measures came from Robert Lane, a political scientist at Yale University. Like most social scientists, Lane found it difficult to keep up with the proliferation of social attitude measures. In the summer of 1958, he attempted to pull together a broad range of scales that would be of interest to researchers in the field of political behavior. Subsequently, this work was continued and expanded at the Survey Research Center of the University of Michigan under the general direction of Philip Converse, with support from a grant by the National Institute of Mental Health.

The result was a three-volume series, the most popular of which was the last, *Measures of Social Psychological Attitudes*. That is the focus of our first update of the original volumes.

Readers will note several differences between this work and its predecessors. Most important, we have given responsibility for each topic to experienced and well-known researchers in each field rather than choosing and evaluating items by ourselves. These experts were also limited to identifying the 10 or 20 most interesting or promising measures in their area, rather than covering all available instruments. This new structure has resulted in more knowledgeable review essays, but at the expense of less standardized evaluations of individual instruments.

There are many reasons for creating a volume such as this. Attitude and personality measures are likely to appear under thousands of book titles, in dozens of social science journals, in seldom circulated dissertations, and in the catalogues of commercial publishers, as well as in undisturbed piles of manuscripts in the offices of social scientists. This is a rather inefficient grapevine for the interested researcher. Too few scholars stay in the same area of study on a continuing basis for several years, so it is difficult to keep up with all of the empirical literature and instruments available. Often, the interdisciplinary investigator is interested in the relation of some new variable, which has come to attention casually, to a favorite area of interest. The job of combing the literature to pick a proper instrument consumes needless hours and often ends in a frustrating decision to forego measuring that characteristic, or worse, it results in a rapid and incomplete attempt to devise a new measure. Our search of the literature has revealed unfortunate replications of previous discoveries as well as lack of attention to better research done in a particular area.
The search procedure used by our authors included thorough reviews of Psychological Abstracts as well as the most likely periodical sources of psychological instruments (e.g., Journal of Personality and Social Psychology, Journal of Personality Assessment, Journal of Social Psychology, Personality and Social Psychology Bulletin, Child Development, and the Journal of Applied Psychology) and sociological and political measures (Social Psychology Quarterly, American Sociological Review, Public Opinion Quarterly, and American Political Science Review). Doctoral dissertations were searched by examining back issues of Dissertation Abstracts. Personal contact with the large variety of empirical research done by colleagues widened the search, as did conversations with researchers at annual meetings of the American Sociological Association and the American Psychological Association, among others. Papers presented at these meetings also served to bring a number of new instruments to our attention.

Our focus in this volume is on attitude and personality scales (i.e., series of items with homogeneous content), scales that are useful in survey or personality research settings as well as in laboratory situations. We have not attempted the larger and perhaps hopeless task of compiling single attitude items, except for ones that have been used in large-scale studies of satisfaction and happiness (see Chapter 3). While these often tap important variables in surveys and experiments, a complete compilation of them (even for happiness) is beyond our means.

Although we have attempted to be as thorough as possible in our search, we make no claim that this volume contains every important scale pertaining to our chapter headings. We do feel, however, that our chapter authors have identified most of the high quality instruments.

Contents of This Volume

A brief outline of the contents of our 12 chapters may prove useful. The remainder of this introductory chapter describes the background of our project and explains the major criteria for scale construction that we asked reviewers to use in evaluating the 150 scales covered in this volume. These evaluative criteria fall into three groups:

1. Item construction criteria (sampling of relevant content, wording of items, and item analysis);
2. Response set criteria (controlling the spurious effects of acquiescence and social desirability response sets); and
3. Psychometric criteria [representative sampling, presentation of normative data, reliability (both test–retest reliability and internal consistency), and validity (both convergent and discriminant)].

Of course, meeting these criteria does not in itself determine the value of a scale. For example, one can construct a scale with high internal consistency merely by including items that express the same idea in much the same way, and one can ensure significant discrimination of known groups merely by sampling two groups so divergent that they would be unlikely to respond the same way to any item. For this reason, we recommend that the choice of a scale from this volume be based as much as possible on decision theoretic criteria, such as those originally outlined by Cronbach and Gleser (1965).

The second chapter contains a review of measures of "response set," the main factor that confounds interpretation of items for researchers in all question domains. Delroy Paulhus describes three major kinds of response sets (social desirability, acquiescence, and extremity) and provides ways of collecting and analyzing attitude data that minimize
1. Criteria for Scale Selection and Evaluation

the effects of these sets. He also reviews seven measures of the major response set, social desirability, and describes the advantages and disadvantages of each.

From this chapter we proceed to the 10 chapters dealing with specific social psychological measures. In Chapter 3, Frank Andrews and John Robinson review 10 scales dealing with the most general affective states: life satisfaction and happiness. Drawing upon Andrews’ experience in conducting many national quality-of-life studies at the Survey Research Center of the University of Michigan, these authors identify background and attitudinal factors related to these affective states. They also describe several single-item measures of satisfaction and happiness as well as some promising new scales designed to measure related constructs.

Chapter 4 contains measures of self-esteem. Jim Blascovich and Joe Tomaka update Shaver’s (1969) and Crandall’s (1973) comprehensive reviews of these measures. They identify the most frequently cited measures in the literature and then review what they consider to be the 11 most worthwhile measures. Five of these are general measures for use with adolescents and adults, two are designed for use with younger children, and the remaining four cover specific aspects of self-esteem.

Chapter 5 reviews 10 measures of social anxiety, conceived as a major inhibitor of social interaction. Mark Leary notes the important distinction between subjective and behavioral manifestations of social anxiety and reviews four subjective measures and six that combine a subjective and a behavioral focus. One of the scales in the latter category has been used in hundreds of studies, although little use has been made of the two factors that are built into the scale. This supports Leary’s general conclusion that the “link between subjective social anxiety and overt behavior is still poorly understood.”

Chapter 6 is devoted to two related negative affective states, depression and loneliness. Phillip Shaver and Kelly Brennan explain similarities and differences between the two states, noting that measures of depression are inextricably tied to official clinical and psychiatric definitions. Both depression and loneliness have been measured in a variety of ways and in an extraordinary number of studies. Shaver and Brennan note that these emotional states have become central issues of our times.

Chapter 7 is concerned with alienation. The 24 scales Melvin Seeman reviews in this chapter cover a wide theoretical terrain and several decades of cumulative research. Seeman outlines separate dimensions of alienation, such as normlessness, powerlessness, and meaninglessness. His scale reviews are then organized in terms of seven of these dimensions and distinctions.

In contrast to these five areas, which directly concern positive and negative affective states, the remaining five chapters deal with beliefs, values, and traits. The first of them, attitudes toward people, is reviewed by Lawrence Wrightsman, who argues in Chapter 8 that these measures have generally been underused, given their strong predictive value in studies of occupational choice and racial attitude change. Work in this area has been hampered by ambiguities in operational definitions of trust. The most widely used and interesting of the eight scales in this chapter is Richard Christie’s Machiavellianism Scale, which is based on astute observations of Italian political life recorded nearly four centuries ago in The Prince.

In Chapter 9 Herbert Lefcourt reviews 16 attitude scales related to locus of control, a term that refers to the presumed internal states that keep some people functioning in difficult periods, while others are overwhelmed with negative emotions. While research interest in this ability to feel that one can control one’s own destiny has waned recently, Lefcourt finds locus of control to be “a stimulating and useful research tool” across a broad range of research. Recent variants of the scale have distinguished among three separate agents of control (internal, chance, and powerful others), different goal states,
and different environments (e.g., personal, interpersonal, sociopolitical, work). Four of the scales Lefcourt reviews deal with general locus of control, three others are for use with children, four deal with specific goal areas, and five deal with health-related content.

Chapter 10 reviews 17 measures related to authoritarianism. In a broad-ranging historical essay, Richard Christie describes the origins of these scales and offers some fascinating insights into their development. He portrays authoritarianism as an under-utilized construct in social psychological research and notes the development of a superior measure in the area, one that transcends the severe methodological limitations of earlier measures. Although factor analytic studies have produced several divergent results, the importance of child-rearing themes in this area has been reinforced in several studies; submission to ingroup figures and institutional control are two further recurrent themes. Christie notes the need for more study with working-class samples and the need for behavioral measures.

In Chapter 11, Ellen Lenney describes eight measures of androgyny. She concludes that the creation of separate, largely orthogonal measures of masculinity and femininity provided an important breakthrough in this line of research. She notes several clear advantages of new measures based on the two-dimensional scheme, in contrast to earlier scales that treated masculinity—femininity as a single dimension.

In the final and most general chapter, Chapter 12, a variety of measures of values are presented. Included in this chapter are instruments based on the work of anthropologists and philosophers, as well as psychologists and sociologists. Valerie Braithwaite and William Scott devote most of their attention to the vast literature generated by Rokeach’s value measures. They also review 14 other value scales, which tap such factors as concern for others, desire for status, individualism, and religiosity, among others. Of particular interest are four scales that measure non-Western orientations toward life. The lack of truly cross-cultural measures and of research relating values and behavior are two of the major shortcomings in this area.

A Broader Conceptual Framework

Stepping back from the specific contents of these 11 chapters, it can be argued that almost all of the measures bear some relation to the five basic emotional factors identified in recent multidimensional research (e.g., Epstein, 1984; Shaver, Schwartz, Kirson, & O’Connor, 1987). When research respondents in the Shaver et al. (1987) study were asked to sort the 135 most prototypical emotions into categories that were similar to one another, five major groupings emerged: love (liking), happiness, anger, sadness (depression), and fear (anxiety).

In the present volume, our chapter headings have been determined primarily by the degree to which various attitude constructs have been cited and reported in the social science literature. It can be seen that love or liking is assessed to some degree by the attitudes toward people scales in Chapter 8 and by the self-esteem (self-love) scales in Chapter 4. Happiness is obviously directly assessed with the measures in Chapter 3, and indirectly with the locus of control scales in Chapter 9. The struggle to maintain a positive affective state such as happiness or self-esteem is reflected in the measures in Chapter 2; social desirability bias can be considered a defensive or self-enhancing strategy. Both masculinity and femininity, discussed in Chapter 11, have been found to be related to measures of self-esteem and well-being.

Considering the hedonically negative emotional categories, anger (hostility, hatred) is related to measures of authoritarianism (Chapter 10) and to cynical or hostile attitudes
toward people (Chapter 8). Sadness is obviously related to depression, loneliness, alienation, (low) self-esteem, and low satisfaction with life, measures of which are reviewed in Chapters 3, 4, 6, and 7. Sadness (depression) is also more common among people with an external locus of control (Chapter 9). Fear and anxiety are directly assessed by the measures of social anxiety in Chapter 5 and are present in people who are high in dogmatism (Chapter 10). Finally, the superordinate categories of positive versus negative emotions are reflected in certain measures of values (Chapter 12).

In this sense, then, the scale topics selected for review in this volume can be seen to relate to basic emotional states. Given the growth of research on emotion in recent years (e.g., Scherer \& Ekman, 1984; Frijda, 1986; Izard, Kagan, \& Zajonc, 1984; and the new journal, Cognition and Emotion), it seems likely that these links between attitudes and basic emotions will become more explicit. Some early evidence of the relation of many of these measures to life satisfaction was reviewed by Robinson (1969).

**Evaluative Criteria**

We have tried to go beyond a simple listing of potential instruments and their psychometric properties. While most scale authors do provide useful statistical data in their scale presentations, it is one thing to present statistical data and another to interpret them. The casual reader or part-time researcher may find it difficult to assess such assets and liabilities when different authors use different statistical procedures. For example, few researchers seem to know that a Guttman reproducibility coefficient of .91 can be obtained from a series of items with inter-item correlation coefficients around .30, or that a test-retest reliability correlation of .50 may indicate a higher reliability than a split-half reliability of .80.

Nor may scale authors be disposed to point out the limitations of their instruments when they are writing articles for publication. Thus, many authors fail to alert readers to their restricted samples, failure to deal with response sets, items that are too complicated for respondents to understand, lack of item analyses, or failure to include certain behaviors and attitudes relevant to the construct at hand. We have tried, where possible, to make such liabilities visible to the reader, although it was simply not feasible with the space and resources available to note all such shortcomings. Originally we had hoped to order the instruments in each chapter according to their probable research value, or to their ability to meet certain desirable standards; that also was not possible. Within each topic area, the instruments we had space to consider often differ so much in purpose or focus that they cannot be arranged along a single quality dimension.

At present, when experienced researchers disagree with our reviewers' assessments, they need to supplement them with their own. We hope that our reviewers have alerted readers to a number of psychometric considerations, not only when deciding which instrument to use, but also in evaluating their own new scales. We have tried to be fair, honest, consistent, and not overly demanding in our evaluations, and we have tried to highlight the merits as well as the limitations of each instrument.

The following brief description of our evaluative criteria proceeds in the general chronological sequence in which attitude instruments are constructed.

**Writing the Items**

The first step for scale builders, and the first dimension on which their work can be evaluated, is writing or locating items to include in a scale. It is usually assumed that the
scale builder knows enough about the field to construct an instrument that will cover an important theoretical construct well enough to be useful to other researchers. If it covers a construct for which instruments are already available, sound improvements over previous measures should be demonstrated.

Three minimal considerations in constructing a scale are:

1. **Proper Sampling of Content:** Proper sampling is not easy to achieve, nor can exact rules be specified for ensuring its achievement (as critics of Louis Guttman's concept of "universe of content" have noted). Nonetheless, one must be aware of the critical role of item sampling procedures in scale construction. Future research may better reveal the population of behaviors, objects, and feelings that ought to be covered in any area, but some examples may suggest ways in which the interested researcher can provide better coverage of a construct domain. Thus investigators of the "authoritarian personality" lifted key sentiments expressed in small group conversations, personal interviews, and written remarks and transformed them into scale items; some of the items consisted of direct verbatim quotations from such materials. In the job satisfaction area, Robinson, Athanasiou, and Head (1967) presented open-ended responses to such questions as "What things do you like best (or don't you like) about your job?" Responses to such questions offer invaluable guidelines to researchers, concerning both the universe of factors to be covered and the weight that should be given to each factor. Other instruments in the job satisfaction area (as elsewhere) were built either on the basis of previous factor analytic work or on responses to questions concerning critically satisfying or dissatisfying situations. Decisions remain to be made about the number of questions needed to cover each factor, but the important first step is to make sure that the main factors have been identified and covered.

2. **Simplicity of Item Wording:** One of the great advantages of obtaining verbatim comments from group discussions or open-ended questions, as people in advertising have discovered, is that such sentiments are usually couched in language easily comprehended and recognized by respondents. Comparing earlier and contemporary instruments, we see that more recently constructed scales contain question wording that is far less stuffy, complex, and esoteric. Even today, however, items developed from college student samples must be edited and adapted for use with more heterogeneous populations. Some helpful advice on these matters is contained in Sudman and Bradburn (1982), Robinson and Meadow (1982), and Converse and Presser (1986). Many other undesirable item-wording practices seem to be going out of style as well: double-barreled items, which contain so many ideas that it is hard to tell why a person agrees or disagrees with them (e.g., "The government should provide low-cost medical care because too many people are in poor health and doctors charge so much money"); items that are so vague they mean all things to all people ("Everybody should receive adequate medical care"); or items that depend on familiarity with little-known facts ("The government should provide for no more medical care than that implied in the Constitution"). Advice about writing items in the negative versus the positive is offered in our discussion of response set.

3. **Item Analysis:** While item wording is something an investigator can manipulate to ensure coverage of intended content, there is no guarantee that respondents will reply to the items in the manner intended by the investigator. Item analysis is one of the most efficient methods for checking whether people are responding to the items in the manner intended. We have encountered several scales whose authors assumed that some a priori division of scale items corresponded to the way their respondents perceived them.

Many methods of item analysis are available, and, in fact, multidimensional analyses (described below under homogeneity, in our discussion of statistical procedures) can be
1. Criteria for Scale Selection and Evaluation

considered the ultimate item analytic procedure. Researchers need not go so far as to factor-analyze their data to select items to be included or discarded, but an item intercorrelation matrix (on perhaps a small subsample or pretest sample) can be a simple and convenient surrogate for determining which items to include, particularly when using most of the statistical packages available for personal computers. If it is hypothesized that five items in a large battery of items (say those numbered 1, 2, 6, 12, and 17) constitute a scale of authoritarianism, then the majority of the 10 inter-item correlations between these five items should be substantial. At the minimum they should be significant at the .05 level. While this minimum may seem liberal, it is in keeping with the degree to which items in the most reputable scales intercorrelate for heterogeneous populations. If items 1, 2, and 17 intercorrelate substantially with each other but item 6 does not correlate well with any of them, then item 6 should be discarded or rewritten. Measuring the degree to which each of the five items correlates with external criteria or outside variables is a more direct device for the selection of items; this may even be preferable to high inter-item correlations. Such item validity approaches provide a built-in validational component for the scale. Wendt (1979), for example, used canonical correlation methods to find that a general alienation scale factored into two distinct scales with different demographic correlates. Exercises using LISREL programs may be similarly useful.

Robinson (1969) reported learning a valuable lesson about the myriad pitfalls in writing items from a simple item analysis of value questions in a national survey. Twelve items had been selected from a previous study that had uncovered four dimensions of value (authoritarianism, expression, individualism, and equalitarianism). One of the individualism items ("It is the man who starts off bravely on his own who excites our admiration") seemed in particular need of reframing for a cross-sectional survey. Accordingly, the item was reworded, "We should all admire a man who starts out bravely on his own." Item analysis revealed that this reformulated item was more closely associated with the three authoritarianism items than with the other two individualism items. Thus, this seemingly innocuous wording change completely altered the value concept tapped by the item.

For researchers who do not have the luxury of pretesting as a way to eliminate or revise unsatisfactory items, the item analysis phase of scale construction can be incorporated into the determination of the dimensionality or homogeneity of the test items. This will ensure that there is empirical as well as theoretical rationale for combining the information contained in various items into a scale.

Avoiding Response Set

A second large area of concern to scale builders is the avoidance of response set. Response set refers to a tendency on the part of individuals to respond to attitude statements for reasons other than the content of the statements. Thus, a person who might want to appear generally agreeable with any attitude statement is said to show an "agreement response set." One defense against response set is to make the scale as interesting and pleasant for the respondent as possible. The more that respondents find the instrument to be dull or unpleasant, the greater the chance that they will not answer carefully or will attempt to finish it as quickly as possible, agreeing indiscriminately or simply checking off the same answer column for each item.

As Delroy Paulhus details in Chapter 2, two major sources of response set need to be controlled:

1. Acquiescence: Most of us have observed people whose attitudes change in accord with the situation. Such people are said to "acquiesce" in anticipation of opposition from others. In the same way, some people are "yeasayers," willing to go along with anything
that sounds good, while others (perhaps optimists) are unwilling to look at the negative side of any issue. These dispositions are reflected in people's responses to attitude questions. Fortunately, it is often possible to separate their "real" attitudes from their tendency to agree or disagree.

There are various levels of attack, all of which involve abandoning a simple agree-disagree or yes-no format. One can first control simple order effects by at least an occasional switching of response alternatives between positive and negative. For simple "yes-no" alternatives, a few "no-yes" options should be inserted. Similarly, for the "strongly agree, agree, uncertain, disagree, strongly disagree" or Likert format, the five alternatives should occasionally be listed in the opposite order. This practice will offer some possibility of locating respondents who choose alternatives solely on the basis of the order in which they appear. It may also encourage overly casual respondents to think more about their answers, although at the cost of some confusion to respondents.

It is more difficult to shift the entire item wording from positive to negative, as those who have tried to reverse authoritarianism items (Chapter 10) have found. A logician may argue that the obverse of "Obedience is an important thing for children to learn" is not "Disobedience is an important thing for children to learn," and the investigator is on shaky ground in assuming that a respondent who agrees with both the first statement and the second is completely confused or vulnerable to agreement response set. Along the same line, the practice of inserting a single word in order to reverse an item can produce rather awkward sentences, while changing one word in an unusual context can produce items in which most respondents may not notice the change. In sum, writing item reversals requires considerable sensitivity and care. The interested researcher should check previous work on the subject (as referenced in Chapter 10).

A third and more difficult, yet probably more effective, approach involves the construction of forced-choice items. Here two (or more) replies to a question are listed and respondents are told to choose only one: "The most important thing for children to learn is (obedience) (independence)." Equating the popularity or social desirability of each of these alternatives provides even greater methodological purity but also entails more intensive effort on the part of both scale constructors and respondents. At the same time, the factor of social desirability is an important response set variable in its own right and needs to be controlled independently of acquiescence.

2. Social Desirability: In contrast to the theory that the acquiescent person reveals a certain desire for subservience in his willingness to go along with any statement, Edwards (1957) proposed more positively that such people are just trying to make a good impression. Decreasing social desirability responding usually involves the use of forced-choice items in which the alternatives have been equated on the basis of social desirability ratings. In more refined instruments, the items are pretested on social desirability, and alternative pairings (or item pairings) that do not prove to be equated are dropped or revised. DeMaio (1984) discusses approaches to the social desirability factor in the context of cross-section surveys.

We have mentioned the major sources of response set contamination, but there are others of which investigators should be aware. One of the more prevalent sources of contamination is the faking of responses according to some preconceived image that the respondent wants to convey. On a job satisfaction scale, for example, the respondent may try to avoid saying anything that might put his supervisor in a bad light or might involve a change in work procedures. College students may be aware of a professor’s hypothesized relationship between two variables and try to answer in ways that confirm (or disconfirm) this prediction. Other undesirable variations of spurious response patterns that an investi-
1. Criteria for Scale Selection and Evaluation

Statistical Criteria

The third area of instrument evaluation concerns the various statistical and psychometric procedures incorporated into its construction. These include respondent sampling, presentation of norms (usually means and standard deviations), reliability, and validity. While each of these statistical considerations is important, inadequate performance on any one of them does not render the scale worthless. Nevertheless, inadequate performance or lack of concern with many of them does indicate that the scale should be used with reservation. Recent scale authors have paid more heed to these considerations than their predecessors did, but few scales can be said to be ideal on all these factors.

The following eight statistical standards cover the basic requirements in the construction of a well-designed scale:

1. Representative Sampling: Too many researchers remain unaware of the fallacy of generalizing results from samples of college students to an older and much less well-educated general population (for an excellent review, see Sears, 1986). Indeed, some statisticians argue that a sample of a single classroom should be treated as a sample size of one, not the number of students in the classroom. Moreover, college students as a whole represent less than 5% of the population of the United States and diverge from the population on two characteristics that survey researchers usually find most distinctive in predicting attitude differences: age and education. Significant differences among college students are also likely to be found between freshmen and seniors, engineering and psychology students, and students at different colleges, so that one must be careful in expecting results from one classroom sample to hold for all college students. In the political attitude area, distinctions made by political elites may not be recognized by typical citizens, or even by politically sophisticated college students.

This is not meant to discourage researchers from improving the representativeness of whatever populations they do have available for study but rather to caution against generalizing from their findings to people not represented by their samples. Nor is it meant to imply that samples of college students are a useless group on which to construct scales. In areas like foreign affairs, one might well argue that college exposure is the best single criterion of whether a person can truly appreciate the intricacies of the issues involved.

However, an instrument constructed from replies of a random cross section of all students in a university has much more to offer than the same instrument developed on students in a single class in psychology (even if there are more students in the classroom than in the university sample). The prime consideration is the applicability of the scale and scale norms to respondents who are likely to use them in the future.

Problems arise with many samples of noncampus respondents as well. Poor sampling frames and low response rates are not uncommon, even for scales that are otherwise carefully designed and administered to community samples.

2. Normative Information: The adequacy of norms (mean scale scores, percentage agreements, etc.) is obviously dependent on the adequacy of the sample. The most basic piece of normative information is the difference between the researcher's sample and the sample on which the scale was developed in terms of mean scale score and standard deviation.

Additional topics of useful statistical information include: item means (or percentage
agreements), standard deviations, and median scores (if the scale scores are skewed). Most helpful are means and standard deviations for certain well-defined groups (e.g., men and women, Catholics and Baptists) who have high or low scale scores. When such differences have been predicted the results bear on the validity of the scale, which is discussed below. Validity, reliability, and homogeneity are also important areas of basic normative information, of course, and they are covered below in more detail.

3. Reliability (test-retest): "Reliability" is one of the most ambiguous terms in psychometrics. There are at least three major referents: (1) the correlation between the same person's scores on the same items at two separate points in time; (2) the correlation between two different sets of items at the same time (called parallel forms if the items are presented in separate formats, and split-half if the items are all presented together); and (3) the correlation between the scale items for all who answer the items. The latter two indices refer to the internal structure or homogeneity of the scale items (the next criterion), while the former indicates stability of a respondent's item responses over time. It is unfortunate that test-retest measures, which require more effort and sophistication on the part of scale authors and may generate lower reliability figures for their efforts, are available for so few instruments. While the test-retest reliability level may be approximately estimated from indices of homogeneity, there is no substitute for the actual test-retest data. Some attempts to assess reliability and stability are discussed in Wheaton et al. (1977) and Bohrnstedt, Mohler, and Muller (1987).

4. Internal Consistency: In addition to split-half, parallel forms, and inter-item indices of the internal homogeneity of scale items, there exist other measures of reliability. Some of these item-test and internal consistency measures have known statistical relationships with one another, as Scott (1960) and others have shown. Even between such "radically" different procedures as the traditional psychometric approach and the Guttman cumulative approach, however, there likely exist reasonably stable relationships between indices based on inter-item, item-total, and total test homogeneity; as yet, however, these have not been charted. This includes the major reliability coefficient, Cronbach's $\alpha$ (1951).

Currently, the major difference between the indices seems to lie in a researcher's preference for large or small numbers. Inter-item correlations and homogeneity indices based on Loevinger's concepts seldom exceed .40. If one prefers larger numbers, a reproducibility coefficient or split-half reliability coefficient computed on the same data could easily exceed .90. While there is currently no way of relating the various indices, one minimal, but still imperfect, criterion is that of statistically significant correlations. Many researchers complain that this criterion depends too heavily on the sample sizes involved. To make the job even more difficult, statistical distributions of these various indices are not always available so that significance can be ascertained.

Of all the proposed indices, none combines simplicity with amount of information conveyed as well as the inter-item correlation matrix. Computing Pearson $r$ correlation coefficients for more than five items is no longer a time-consuming operation for any researcher with access to a personal computer. Even the inter-item correlation matrix for a 20-item scale can now be generated in a matter of seconds. In the case of dichotomous (two-choice) items, the coefficient Yule's $Y$ or Kendall's tau-B can easily be calculated to determine inter-item significance. Cronbach's $\alpha$ is now calculated on personal computer scaling programs. These, however, constitute only rule-of-thumb procedures for deciding whether a group of items should be added together to form a scale or index. Similarly, the criterion of statistical significance is proposed only because it is a standard that remains fairly constant across the myriad measures which are now, or have been, in vogue. Perhaps more satisfactory norms may be proposed in the future.
1. Criteria for Scale Selection and Evaluation

When the number of items goes beyond 10, however, the inter-item matrix becomes quite cumbersome to analyze by inspection. One is well advised to have the data analyzed by a multidimensional computer program. Program packages such as SPSS and SAS have the ability to factor-analyze 10–50 item intercorrelations in a few minutes, given a reasonably sized sample. These sorts of analyses will help one locate groups of items that go together much faster than could be done by inspecting the correlation matrix.¹ There are many kinds of factor analysis programs and options; under most circumstances, however, the differences between them usually do not result in radically different factor structures.

To say that factor analytic programs do not usually vary greatly in their output is not to imply that structures uncovered by factor analysis are without serious ambiguities. In particular, one common structure of attitudinal data seems to produce an indeterminant factor structure. This occurs when almost all the items are correlated in the range from about .15 to .45. Sometimes only a single factor will emerge from such a matrix and sometimes a solution will be generated that more clearly reflects item differentiation on a series of factors. We have encountered one instance in which an instrument that was carefully constructed to reflect a single dimension of inner-versus other-directedness (according to a forced-choice response format) was found to contain eight factors when presented in Likert format. Thus, one can offer no guarantee that building scales based on inter-item significance will invariably generate unidimensional scales. Nonetheless, only by these procedures can scale authors properly separate the apples, oranges, and coconuts in the fruit salad of items they have assembled.

One final word of caution: It is possible to devise a scale with very high internal consistency merely by writing the same item in a number of different ways. Obviously, such scales tap an extremely narrow construct. Sampling of item content, therefore, is crucial in assessing internal consistency. Internal consistency is a very desirable property, but it needs to be balanced by concept coverage, proper norms, and careful validity work.

5. Known Groups Validity: Validity is the more crucial indicator of the value of the scale. Nevertheless, group discrimination is not necessarily the most challenging hurdle to demonstrated validity. It is rather difficult to construct a liberalism–conservatism scale that will not show significant differences between members of The Heritage Foundation and members of the American Civil Liberties Union, or a religious attitude scale that will not separate Mormons from Jews or ministerial students from engineers. The more demanding criterion is whether the scale scores reliably distinguish happy from miserable people, liberals from conservatives, agnostics from believers within heterogeneous samples—or predict which of them will demonstrate behavior congruent with their attitudes.

6. Convergent Validity (Predictions from Theory): A second and more usual test of convergent validity involves obtaining results from the scale consistent with one’s theory. For example, one might find that older people or better educated people or students with higher grades score higher on the scale, which would be consistent or convergent with some theoretical expectation or prediction. One might also expect that the scale scores would be higher among people who engaged in some type of behavior (such as joining a social group or contributing money) or expressed a particular attitude. The persuasiveness of this convergent or construct validation depends of course on the comprehensiveness and plausibility of the theory and the strength of the outside correlations. More formal attempts to establish construct validity have been attempted through causal modeling (e.g., Andrews, 1984).

¹Researchers should not be deceived by what appear to be high factor loadings. Factor loadings need to be squared to reach levels that are equivalent to correlation coefficients.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Theoretical development/structure</td>
<td>Reflects several important works in the field plus extensive face validity check</td>
<td>Either reviews several works or extensive face validity</td>
<td>Reviews more than one source</td>
<td>Reviews one (no sources)</td>
<td>Ad hoc</td>
</tr>
<tr>
<td>Pilot testing/item development</td>
<td>More than 250 items in the initial pool; several pilot studies</td>
<td>100–250 items in initial pool; more than two pilot studies</td>
<td>50–100 items in initial pool; two pilot studies</td>
<td>Some items eliminated; one small pilot study</td>
<td>All initial items included; no pilot study</td>
</tr>
<tr>
<td>Available norms</td>
<td>Means and SDs for several subsamples and total sample; extensive information for each item</td>
<td>Means and SDs for total and some groups; some item information</td>
<td>Means for some subgroups; information for some items</td>
<td>Means for total group only; information for 1–2 items</td>
<td>None; no item information</td>
</tr>
<tr>
<td>Samples of respondents</td>
<td>Random sample of nation/community with response rate over 60%</td>
<td>Cross-sectional sample of nation/community; random national sample of college students</td>
<td>Some representation of non-college groups; random sample of college students in same departments or colleges</td>
<td>Two or more college classes (some heterogeneity)</td>
<td>One classroom group only (no heterogeneity)</td>
</tr>
<tr>
<td>Inter-item correlations</td>
<td>Inter-item correlation average of .30 or better</td>
<td>Inter-item correlation average of .20-.29</td>
<td>Inter-item correlation average of .10-.19</td>
<td>Inter-item correlations below .10</td>
<td>No inter-item analysis reported</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----------------------------------------------</td>
<td>------------------------------------------</td>
<td>------------------------------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>Coefficient α</td>
<td>.80 or better</td>
<td>.70-.79</td>
<td>.60-.69</td>
<td>&lt;.60</td>
<td>Not reported</td>
</tr>
<tr>
<td>Factor analysis</td>
<td>Single factor from factor analysis</td>
<td>Single factor from factor analysis</td>
<td>Single factor from factor analysis</td>
<td>Some items on same factors</td>
<td>No factor structure</td>
</tr>
<tr>
<td>Test–retest</td>
<td>Scale scores correlate more than .50 across at least a 1-year period</td>
<td>Scale scores correlate more than .40 across a 3–12-month period</td>
<td>Scale scores correlate more than .30 across a 1–3-month period</td>
<td>Scale scores correlate more than .20 across less than a 1-month period</td>
<td>No data reported</td>
</tr>
<tr>
<td>Known groups validity</td>
<td>Discriminate between known groups highly significantly; groups also diverse</td>
<td>Discriminate between known groups highly significantly</td>
<td>Discriminate between known groups significantly</td>
<td>Discriminate between known groups</td>
<td>No known groups data</td>
</tr>
<tr>
<td>Convergent validity</td>
<td>Highly significant correlations with more than two related measures</td>
<td>Significant correlations with more than two related measures</td>
<td>Significant correlations with two related measures</td>
<td>Significant correlation with one related measure</td>
<td>No significant correlations reported</td>
</tr>
<tr>
<td>Discriminant validity</td>
<td>Significantly different from four or more unrelated measures</td>
<td>Significantly different from two or three unrelated measures</td>
<td>Significantly different from one unrelated measure</td>
<td>Different from one correlated measure</td>
<td>No difference or no data</td>
</tr>
<tr>
<td>Freedom from response set</td>
<td>Three or more studies show independence</td>
<td>Two studies show independence</td>
<td>One study shows independence</td>
<td>Some show independence, others do not</td>
<td>No tests of independence</td>
</tr>
</tbody>
</table>
7. **Cross-Validation**: Cross-validation requires two different samples and measures of some criterion variable on each sample. The question to be answered by the test is whether the combination of items for sample A that best correlates with the criterion variable will also work for sample B, and conversely, whether the best set of sample B items works for sample A. Note that the crux of the procedure involves first identifying the items from sample A and then testing them independently on sample B.

8. **Discriminant Validation**: A more refined and powerful standard is the multitrait-multimethod matrix as proposed by Campbell and Fiske (1959). The method requires more than one index of each of the several constructs (say x, y, and z) one wants to measure with the instrument. It is best to include as many measures or indices of each construct as possible, as well as to measure for control purposes such variables as intelligence or response set that could also explain apparent relationships. In the resulting correlation matrix, the various items measuring each single construct (say x) should first correlate highly among themselves; second, the correlations among these items should be higher than their correlations with the items intended to measure constructs y or z, or any of the control variables. The latter is evidence of the scale's ability to discriminate from these other measures.

Needless to say, this is a gross oversimplification of the Campbell–Fiske criteria, and interested readers should peruse the authors' article thoroughly before attempting comparable analyses. At the time they were writing, Campbell and Fiske found only a few personality scales which met their conditions. One recent example of an attitude scale that meets them is Andrews and Crandall's (1976) life quality scale.

In certain chapters in our earlier volumes, in which a sufficient number of instruments to warrant comparison was present, we attempted to rate each scale on all such considerations. A current rating scheme is shown in Table 1.

 Readers might consider using this scheme to evaluate the adequacy of measures they propose to use. Readers are also referred to the psychometric standards of the American Psychological Association (1985) and those in Heise and Bohrnstedt (1970), Bohrnstedt and Borgatta (1981), and Werts and Linn (1970), for "quick methods" of ascertaining reliability and validity.

 Even this extensive list of proposed criteria is far from exhaustive. The actual choice of an instrument should be dictated by decision theoretic considerations. Thus the increasing of homogeneity by adding questionnaire items needs to be balanced against corresponding increases in administrative analysis and cost (as well as respondent fatigue and noncooperation) before one decides on how many attitude items to use. For assessing general levels of some attitude (e.g., separating believers from atheists), well-worded single items may do the job just as well as longer scales no matter how competently the scales are devised.

**Future Volumes**

As noted at the outset, this is the first of a planned series of updated measurement volumes. Subsequent volumes in this series will deal with (a) measures of role-related attitudes (e.g., toward work, marriage, parenthood) and (b) politically relevant scales (as in our Measures of Political Attitudes). Areas for further volumes are being considered, and further revisions of the present volume will appear as demand and changes in the field warrant.

