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 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 investigator may wish to minimize can result from the respondents' wanting (a) to appear too consistent, (b) to use few or many categories in their replies, or (c) to choose extreme alternatives.

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 α (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 α 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.

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).

[1] Researchers 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 1

Some General Rating Criteria for Evaluating Attitude Measures

Criterion rating	4. Exemplary	3. Extensive	2. Moderate	1. Minimal	0. None
Theoretical development/structure	Reflects several important works in the field plus extensive face validity check	Either reviews several works or extensive face validity	Reviews more than one source	Reviews one (no sources)	Ad hoc
Pilot testing/item development	More than 250 items in the initial pool; several pilot studies	100–250 items in initial pool; more than two pilot studies	50–100 items in initial pool; two pilot studies	Some items eliminated; one small pilot study	All initial items included; no pilot study
Available norms	Means and SDs for several subsamples and total sample; extensive information for each item	Means and SDs for total and some groups; some item information	Means for some subgroups; information for some items	Means for total group only; information for 1–2 items	None; no item information
Samples of respondents	Random sample of nation/community with response rate over 60%	Cross-sectional sample of nation/community; random national sample of college students	Some representation of non-college groups; random sample of college students in same departments or colleges	Two or more college classes (some heterogeneity)	One classroom group only (no heterogeneity)

Inter-item correlations	Inter-item correlation average of .30 or better	Inter-item correlation average of .20–.29	Inter-item correlation average of .10–.19	Inter-item correlations below .10	No inter-item analysis reported
Coefficient α	.80 or better	.70–.79	.60–.69	<.60	Not reported
Factor analysis	Single factor from factor analysis	Single factor from factor analysis	Single factor from factor analysis	Some items on same factors	No factor structure
Test–retest	Scale scores correlate more than .50 across at least a 1-year period	Scale scores correlate more than .40 across a 3–12-month period	Scale scores correlate more than .30 across a 1–3-month period	Scale scores correlate more than .20 across less than a 1-month period	No data reported
Known groups validity	Discriminate between known groups highly significantly; groups also diverse	Discriminate between known groups highly significantly	Discriminate between known groups significantly	Discriminate between known groups	No known groups data
Convergent validity	Highly significant correlations with more than two related measures	Significant correlations with more than two related measures	Significant correlations with two related measures	Significant correlation with one related measure	No significant correlations reported
Discriminant validity	Significantly different from four or more unrelated measures	Significantly different from two or three unrelated measures	Significantly different from one unrelated measure	Different from one correlated measure	No difference or no data
Freedom from response set	Three or more studies show independence	Two studies show independence	One study shows independence	Some show independence, others do not	No tests of independence