Psychological Assessment
1995, Vol. 7, No. 3, 309-319


1040-3590/95/$3.00

# Constructing Validity: Basic Issues in Objective Scale Development

Lee Anna Clark and David Watson
The University of Iowa

A primary goal of scale development is to create a valid measure of an underlying construct. We discuss theoretical principles, practical issues, and pragmatic decisions to help developers maximize the construct validity of scales and subscales. First, it is essential to begin with a clear conceptualization of the target construct. Moreover, the content of the initial item pool should be overinclusive and item wording needs careful attention. Next, the item pool should be tested, along with variables that assess closely related constructs, on a heterogeneous sample representing the entire range of the target population. Finally, in selecting scale items, the goal is unidimensionality rather than internal consistency; this means that virtually all interitem correlations should be moderate in magnitude. Factor analysis can play a crucial role in ensuring the unidimensionality and discriminant validity of scales.

Scale development remains a growth industry within psychology. A PsycLIT database survey of articles published in the 6-year period from 1989 through 1994 revealed 1,726 articles with the key words "test construction" or "scale development" published in English-language journals, 270 in other-language journals, and 552 doctoral dissertations. During this same period (i.e., beginning with its inception), 50 articles addressing scale development or test construction were published in *Psychological Assessment* alone. The majority of these articles reported the development of one or more new measures (82%); most of the rest presented new scales derived from an existing instrument (10%). We use these 41 scale-development articles as a reference set for our discussion. Clearly, despite the criticism leveled at psychological testing in recent years, assessment retains a central role within the field.

Given that test construction remains a thriving activity, it is worthwhile to reconsider the scale development process periodically to maintain and enhance the quality of this enterprise. The goal of this article is to articulate some basic principles that we believe anyone developing a scale should know and follow. Many of these principles have been stated before, but we repeat them here both because they are sufficiently important to bear repetition and because a review of the recent literature indicates that they are still not universally honored.

We focus on verbally mediated measures; thus, for example, we do not address the development of behavioral observation scales. Moreover, our primary focus is on self-report measures, because these constitute the majority (67%) of our reference sample. Nonetheless, most of the basic principles we articulate are applicable to interview-based measures and rating scales designed to be completed by clinicians, parents, teachers, spouses, peers, and so forth.

Before proceeding further, it is interesting to examine the new measures comprising our *Psychological Assessment* sample. This examination sample offers a glimpse at why scale development continues unabated, as well as the nature of the unmet needs these scale developers are seeking to fill. First, not surprisingly given this journal's focus, more than half (61%) of the scales assess some aspect of psychopathology, personality, or adjustment. The next most common categories are measures of attitudes and interpersonal relations (20% and 15%, respectively). The remaining scales assess a miscellany of behaviors, abilities, response validity, trauma experience, and so forth. In all categories, most new scales apparently tap relatively narrow constructs, such as suicidality, fear of intimacy, postpartum adjustment, drug-use expectancies, or parent–teenager relations, that have a focused range of utility. However, the extent to which the score variance of such scales is, in fact, attributable to the named target construct is an important issue that we will consider.

## The Centrality of Psychological Measurement

It has become axiomatic that (publishable) assessment instruments are supposed to be reliable and valid; indeed, every article in the *Psychological Assessment* set addresses these qualities. However, it appears that many test developers do not fully appreciate the complexity of these concepts. As this article is being prepared, the *Standards for Educational and Psychological Testing* (American Psychological Association, 1985) are undergoing intensive review and revision for the first time in a decade. Strong and conflicting pressures regarding the *Standards'* revision are being brought to bear on the Joint Committee on the Standards for Educational and Psychological Testing by diverse groups, and major changes in the *Standards* are expected. Whatever else it may do, however, the Joint Committee intends to emphasize the centrality of construct validity in testing even more than in previous versions, according to Co-Chair C. D. Spielberger (personal communication, February 15, 1995). And yet, widespread misunderstanding remains regarding pre-

Lee Anna Clark and David Watson, Department of Psychology, The University of Iowa.

We wish to acknowledge the helpful comments of Jane Loevinger on a previous version of this article.

Correspondence concerning this article should be addressed to Lee Anna Clark, Department of Psychology, The University of Iowa, E11 Seashore Hall, Iowa City, Iowa 52242-1407. Electronic mail may be sent via Internet to laclark@blue.weeg.uiowa.edu.

cisely what construct validity is and what establishing construct validity entails.

Cronbach and Meehl (1955) argued that investigating the construct validity of a measure necessarily involves at least the following three steps: (a) articulating a set of theoretical concepts and their interrelations, (b) developing ways to measure the hypothetical constructs proposed by the theory, and (c) empirically testing the hypothesized relations among constructs and their observable manifestations. This means that without an articulated theory (which Cronbach and Meehl termed "the nomological net"), there is no construct validity. The Joint Committee's emphasis on the centrality of construct validity is therefore highly appropriate because the process of establishing construct validity represents a key element in differentiating psychology as a science from other, nonscientific approaches to the analysis of human behavior.

Construct validity cannot be inferred from a single set of observations, whether these pertain to a measure's factor structure, correlations with other measures, differentiation between selected groups, or hypothesized changes over time or in response to an experimental manipulation. Clearly, a series of investigations is required even to begin the process of identifying the psychological construct that underlies a measure. Nonetheless, Cronbach and Meehl's (1955) dictum that "One does not validate a test, but only a principle for making inferences" (p. 297) is often ignored, as scale developers speak lightly— sometimes in a single sentence—of establishing the construct validity of a scale. Even the more straightforward concept of reliability is widely mistreated, as we discuss in a later section.

It also should be noted that construct validity is important from the standpoint of practical utility as well as science. That is, for economic reasons, practitioners increasingly are being asked to justify the use of specific assessment procedures to third-party payers. Clear documentation of the precision and efficiency of psychological measures will be required in the near future. The most precise and efficient measures are those with established construct validity; they are manifestations of constructs in an articulated theory that is well supported by empirical data. Thus, construct validity lies at the heart of the clinical utility of assessment and should be respected by scale developers and users alike.

## A Theoretical Model for Scale Development

Loevinger's (1957) monograph arguably remains the most complete exposition of theoretically based psychological test construction. Like any great work, however, her monograph requires exegesis, and in this article we assume this role. Specifically, we offer practical guidance for applying Loevinger's theoretical approach to the actual process of scale development. We limit ourselves to that portion of her article that details the "three components of construct validity," which she labels *substantive, structural,* and *external.* More specifically, because our topic is initial scale development, we focus primarily on the first two of these components, which together address a measure's "internal validity" (Loevinger, 1957, p. 654). Smith and McCarthy's (1995) article in this special issue addresses the external component more thoroughly.

### Substantive Validity: Conceptualization and Development of an Initial Item Pool

#### *Conceptualization*

Our PsycLIT database search suggests that human psychology is sufficiently complex that there is no limit to the number of psychological constructs that can be operationalized as scales. One now widely recognized reason for this is that psychological constructs are ordered hierarchically at different levels of abstraction or breadth (see Comrey, 1988; John, 1990; Watson, Clark, & Harkness, 1994). In the area of personality, for example, one can conceive of the narrow traits of talkativeness and physical expressiveness, the somewhat broader concepts of gregariousness and assertiveness, and the still more general disposition of extraversion. Scales can be developed to assess constructs at each of many levels of abstraction. Consequently, a key issue to be resolved in the initial developmental stage is the scope or generality of the target construct.

As mentioned, our *Psychological Assessment* sample consists primarily of scales that assess narrow-band (e.g., Cocaine Expectancy Questionnaire; Jaffe & Kilbey, 1994) or midlevel (Social Phobia and Anxiety Inventory; Turner, Beidel, Dancu, & Stanley, 1989) constructs. It is noteworthy, therefore, that Loevinger (1957) argued that, even when relatively narrow measurements are desired, those scales based on a "deeper knowledge of psychological theory" (p. 641) will be more helpful in making specific pragmatic decisions than those developed using a purely "answer-based" technology. Accordingly, even narrow-band measures should be embedded in a theoretical framework, and even measures of the same basic phenomenon will vary with the theoretical perspective of the developer.

A critical first step is to develop a precise and detailed conception of the target construct and its theoretical context. We have found that writing out a brief, formal description of the construct is very useful in crystallizing one's conceptual model. For example, in developing the Exhibitionism scale of the Schedule for Nonadaptive and Adaptive Personality (SNAP; Clark, 1993), the initial target construct was defined as a continuum ranging from normal adaptive functioning to potentially pathological behavior of which the high end was defined by overly dramatic, reactive, and intensely expressed behavior; an exaggerated expression of emotions; excessive attention-seeking behavior; an inordinate need for admiration; vanity; and a demanding interpersonal style.

This emphasis on theory is not meant to be intimidating. That is, we do not mean to imply that one must have a fully articulated set of interrelated theoretical concepts before embarking on scale development. Our point, rather, is that thinking about these theoretical issues prior to the actual process of scale construction increases the likelihood that the resulting scale will make a substantial contribution to the psychological literature.

#### *Literature Review*

To articulate the basic construct as clearly and thoroughly as possible, it is necessary to review the relevant literature to see how others have approached the same problem. Initially, the review should include previous attempts to conceptualize and assess both the same construct and closely related constructs.

For instance, in developing a new measure of hopelessness, a thorough literature search would encompass measures of related constructs at various levels of the hierarchy in which the target construct is embedded—for example, depression and optimism–pessimism—in addition to existing measures of hopelessness.

Subsequently, the review should be broadened to encompass what may appear to be less immediately related constructs to articulate the conceptual boundaries of the target construct. That is, in the initial stages one investigates existing scales and concepts to which the target is expected to be related. Then, one also must examine entities from which the target is to be distinguished. In other words, a good theory articulates not only what a construct is, but also what it is not. Continuing with the hopelessness example, a thorough review would reveal that various measures of negative affect (depression, anxiety, hostility, guilt and shame, dissatisfaction, etc.) are strongly intercorrelated, so that it is important to articulate the hypothesized relation of hopelessness to other negative affects. Similarly, a good measure will have a predicted convergent and discriminant correlational pattern (Smith & McCarthy, 1995), and it is important to consider this aspect of measurement at the initial as well as later stages of development.

The importance of a comprehensive literature review cannot be overstated. First, such a review will serve to clarify the nature and range of the content of the target construct. Second, a literature review may help to identify problems with existing measures (e.g., unclear instructions or problematic response formats) that then can be avoided in one's own scale. Finally, and perhaps most importantly, a thorough review will indicate whether the proposed scale is actually needed. If reasonably good measures of the target construct already exist, why create another? Unless the prospective test developer can clearly articulate ways in which the proposed scale will represent either a theoretical or an empirical improvement over existing measures, it is preferable to avoid contributing to the needless proliferation of assessment instruments.

### *Creation of an Item Pool*

Once the scope and range of the content domain have been tentatively identified, the actual task of item writing can begin. No existing data-analytic technique can remedy serious deficiencies in an item pool. Accordingly, the creation of the initial pool is a crucial stage in scale construction. The fundamental goal at this stage is to sample systematically all content that is potentially relevant to the target construct. Loevinger (1957) offered the classic articulation of this principle: *"The items of the pool should be chosen so as to sample all possible contents which might comprise the putative trait according to all known alternative theories of the trait"* (p. 659, emphasis in original).

Two key implications of this principle are that the initial pool (a) should be broader and more comprehensive than one's own theoretical view of the target construct and (b) should include content that ultimately will be shown to be tangential or even unrelated to the core construct. The logic underlying this principle is simple: Subsequent psychometric analyses can identify weak, unrelated items that should be dropped from the emerging scale but are powerless to detect content that should have been included but was not. Accordingly, in creating the item pool one always should err on the side of overinclusiveness. The importance of the initial literature review becomes quite obvious in this connection.

In addition to sampling a sufficient breadth of content, the scale developer must ensure that there is an adequate sample of items within each of the major content areas comprising the broadly conceptualized domain; failure to do so may mean that one or more of these areas will be underrepresented in the final scale. To ensure that each important aspect of the construct is assessed adequately, some test developers have recommended that formal subscales be created to assess each major content area. Hogan (1983), for instance, identified 10 content areas (e.g., anxiety, guilt, and somatic complaints) that make up the more general dimension of Adjustment versus Maladjustment and created 4- to 10-item "homogeneous item composites" to assess each of them. Similarly, Comrey (1988) has championed the use of "factored homogeneous item dimensions" to assess individual content areas within a specified domain.

The important point here is not that a particular procedure must be followed, but that scale developers need to ensure that each content area is well represented in the initial item pool. If only one or two items are written to cover a particular content area, then the chances of that content being represented in the final scale are much reduced. Loevinger (1957) recommended that the proportion of items devoted to each content area be proportional to the importance of that content in the target construct. This is a worthy goal, although in most cases the theoretically ideal proportions will be unknown. However, broader content areas should probably be represented by more items than narrower content areas.

Many of the procedures that we are discussing are traditionally described as the *theoretical–rational* or *deductive* method of scale development. We consider this approach to be an important initial step in a more extensive process rather than a scale development method to be used by itself. Similarly, Loevinger (1957) affirmed that content issues must always be considered in defining the domain, but emphasized that alone they are insufficient. That is, empirical validation of content (as distinguished from "blind empiricism") is important: "If theory is fully to profit from test construction . . . every item [on a scale] must be accounted for" (Loevinger, 1957, p. 657). This obviously is a very lofty goal and clearly is articulated as an ideal to be striven for rather than an absolute requirement (for a very similar view, see Comrey, 1988). For further discussion of content validity issues, see Haynes, Richard, and Kubany (1995) in this special issue.

In this context, we emphasize that good scale construction typically is an iterative process involving several periods of item writing, followed in each case by conceptual and psychometric analysis. These analyses serve to sharpen one's understanding of the nature and structure of the target domain as well as to identify deficiencies in the initial item pool. For instance, a factor analysis might establish that the items can be subdivided into several subscales but that the initial pool does not contain enough items to assess each of these content domains reliably. Accordingly, new items need to be written and again subjected to psychometric analyses. Alternatively, analyses may suggest that conceptualization of the target construct as, for example, a single bipolar dimension is countermanded by evidence that the

two poles actually represent separate and distinct entities. In this case, revision of one's theoretical model may be in order.

An examination of the *Psychological Assessment* sample of scale development articles indicates that most test developers did start with a large item pool that was reduced to a smaller final set. However, it is not clear whether this finding reflects the broad and systematic domain sampling that we advocate or, alternatively, the mere elimination of items that were psychometrically weak for any number of reasons. That is, we saw little evidence of an iterative process through which the conceptualization of the target construct was itself affected by the process of scale development (see Smith & McCarthy, 1995, and Tellegen & Waller, in press, for discussions of this issue).

*Basic principles of item writing.* In addition to sampling well, it also is essential to write "good" items. When developing a scale it is worth the time to consult the available literature on item writing (e.g., Angleitner & Wiggins, 1985; Comrey, 1988; Kline, 1986). What constitutes a good item? First, the language should be simple, straightforward, and appropriate for the reading level of the scale's target population. For instance, scales intended for use in general clinical samples need to be readily understandable by respondents with only a modest education. In addition, one should avoid using trendy expressions that quickly may become dated, as well as colloquialisms and other language for which the familiarity (and thus utility) will vary widely with age, ethnicity, region, gender, and so forth. Finally, there is little point in writing items that virtually everyone (e.g., "Sometimes I am happier than at other times") or no one (e.g., "I am always furious") will endorse, unless they are intended to assess invalid responding. For this and other reasons we discuss later, items should be written to ensure variability in responding.

Item writers also should be careful to avoid complex or "double-barreled" items that actually assess more than one characteristic. At best, such items are ambiguous; at worst, they may leave respondents with no viable response alternative. Consider, for example, the true–false item, "I would never drink and drive for fear that I might be stopped by the police," which confounds the occurrence versus nonoccurrence of a behavior (drinking and driving) with a putative motive for that behavior (fear of legal complications). As such, it may leave respondents who avoid drinking and driving—but who do so for other reasons (e.g., because it is dangerous or morally wrong)—puzzled as to how they should respond. Of equal or greater concern is the fact that respondents will interpret complex items in different ways; accordingly, their responses will reflect the heterogeneity of their interpretations, and the item likely will show very poor psychometric properties as a result.

Furthermore, the exact phrasing of items can exert a profound influence on the construct that is actually measured. This is well illustrated by the example of the general personality trait of neuroticism (negative affectivity; Watson & Clark 1984). Over the years, it has been demonstrated repeatedly that attempts to assess a specific construct (such as hardiness or pessimism) have yielded instead yet another measure that is strongly saturated with this pervasive dimension. Indeed, items must be worded very carefully to avoid tapping into the broad individual differences in affect and cognition that characterize neuroticism. For instance, our own experience has shown that the inclusion of almost any negative mood term (e.g., "I worry about . . .," or "I am upset [or bothered or troubled] by . . .") virtually guarantees that an item will have a substantial neuroticism component; the inclusion of several such affect-laden items, in turn, ensures that the resulting scale—regardless of its intended construct—will be primarily a marker of neuroticism.

*Choice of format.* Finally, in creating the initial item pool, the test developer also must decide on the response format to be used. Clearly, the two dominant response formats in contemporary personality assessment are dichotomous responding (e.g., true–false and yes–no) and Likert-type rating scales with three or more options. Checklists, forced-choice, and visual analog measures also have been used over the years, but for various reasons have fallen out of favor. *Checklists*—scales that permit respondents to scan a list and check only the applicable items—proved to be problematic because they are more prone to response biases than formats that require a response to every item (Bentler, 1969; D. P. Green, Goldman, & Salovey, 1993). Most *forced-choice* formats, in which respondents must choose between alternatives that represent different constructs, are limited in that the resulting scores are ipsative; that is, they reflect only the relative intraindividual strength of the assessed constructs and do not provide normative, interindividual information. Finally, *visual analog* scales provide a free range of response options along a defined continuum, usually anchored at the two endpoints (e.g., *No pain at all* vs. *Excruciating pain; worst I can imagine*). This scale type is rarely used for multi-item scales because they are extremely laborious to score, although this may change with increased use of computer administration. Thus, they are most useful when a single (or few) measurements are desired and the target construct is either very simple (e.g., a single mood term) or represents a summary judgment (e.g., bodily pain).[1]

There are several considerations in choosing between dichotomous and Likert-type formats; furthermore, in the latter case, one also has to decide the number of response options to offer and how to label the response options. Comrey (1988) has criticized dichotomous response formats extensively, arguing that "multiple-choice item formats are more reliable, give more stable results, and produce better scales" (p. 758). Comrey's points are cogent and should be taken very seriously, especially his valid assertion that dichotomous items with extremely unbalanced response distributions (i.e., those in which virtually everyone answers either true or false) can lead to distorted correlational results. However, this problem can be avoided by carefully inspecting individual item frequencies during scale development and eliminating items with extreme response rates (one often-used cutoff is any item on which more than 95% of all respondents give the same response). Furthermore, dichotomous response formats offer an important advantage over rating scales: Other things being equal, respondents can answer many more items in the same amount of time. Consequently, if assessment time is limited, dichotomous formats can yield significantly more information. Moreover, Loevinger (1957) has argued that response biases are more problematic with Likert-type scales and that the assumption of equal-interval scaling often is not justified.

---

[1] We are grateful to an anonymous reviewer for providing additional information regarding visual analog scales.

Likert-type scales are used with a number of different response formats; among the most popular are the frequency (*never* to *always*), degree or extent (*not at all* to *very much*), similarity (*like me* to *not like me*), and agreement (*strongly agree* to *strongly disagree*) formats. Obviously, the nature of the response option constrains item content in an important way (see Comrey, 1988). For example, the item "I often lose my temper" would be inappropriate if used with a frequency format. Note also that with an odd number of response options (typically, five or seven), the label for the middle option must be considered carefully; for example, *cannot say* confounds possible uncertainty about item meaning with a midrange rating of the attribute. An even number of response options (typically, four or six) eliminates this problem but forces respondents to "fall on one side of the fence or the other," which some may find objectionable. In a related vein, it must be emphasized also that providing more response alternatives (e.g., a 9-point rather than a 5-point scale) does not necessarily enhance reliability or validity. In fact, increasing the number of alternatives actually may reduce validity if respondents are unable to make the more subtle distinctions that are required. That is, having too many alternatives can introduce an element of random responding that renders scores less valid.

Finally, we emphasize that dichotomous and rating scale formats typically yield very similar results. For example, neuroticism scales using various formats (including true–false, yes–no, and rating scales) all are highly intercorrelated and clearly define a single common factor (Watson, Clark, & Harkness, 1994). In light of these considerations, we cannot conclude that one type of format is generally preferable to the other. Used intelligently, both formats can yield highly reliable and valid scales. To ensure such intelligent usage, we strongly recommend that a proposed format be pilot-tested on a moderately sized sample to obtain preliminary information about both respondent reactions and response option distributions.

## Structural Validity: Item Selection and Psychometric Evaluation

### *Test Construction Strategies*

The choice of a primary test construction or item selection strategy is as important as the compilation of the initial item pool. In particular, the item selection strategy should be matched to the goal of scale development and to the theoretical conceptualization of the target construct. In this regard, Loevinger (1957) described three main conceptual models: (a) quantitative (dimensional) models that differentiate individuals with respect to degree or level of the target construct, (b) class models that seek to categorize individuals into qualitatively different groups, and (c) more complex dynamic models.

It is beyond the scope of this article to discuss either dynamic or class models; however, we note with concern that some of the articles in the *Psychological Assessment* sample applied methods more appropriate for quantitative models (e.g., factor analysis) to constructs that appeared to reflect class models (such as diagnoses). Of course, some theoreticians have argued that the empirical data do not strongly support class models even in the case of psychiatric diagnoses (e.g., Clark, Watson, & Reynolds, 1995) and, therefore, that dimensional or quantitative models are more appropriate. Thus, these aforementioned *Psychological Assessment* scale developers may have implicitly accepted this stance in selecting their test construction method. In any case, analytic methods appropriate for class model constructs do exist and should be used to develop measures of such constructs (e.g., Gangestad & Snyder, 1991; Meehl & Golden, 1982).

Loevinger (1957) advanced the concept of structural validity, that is, the extent to which a scale's internal structure (i.e., the interitem correlations) parallels the external structure of the target trait (i.e., correlations among nontest manifestations of the trait). She also emphasized that items should reflect the underlying (latent) trait variance. These three concerns parallel the three main item selection strategies in use for quantitative model constructs: empirical (primarily reflecting concern with nontest manifestations), internal consistency (concerned with the interitem structure), and item response theory (focused on the latent trait). The fact that structural validity encompasses all three concerns demonstrates that these methods may be used in conjunction with one another and that exclusive reliance on a single method is neither required nor necessarily desirable.

*Criterion-based methods.* Meehl's (1945) "empirical manifesto" ushered in the heyday of empirically keyed test construction. Backed by Meehl's cogent arguments that a test response could be considered verbal behavior in its own right—with nontest correlates to be discovered empirically—test developers embraced criterion keying as a method that permitted a wide range of practical problems to be addressed in an apparently straightforward manner. With widespread use, however, the limitations of this approach quickly became evident. From a technical viewpoint, major difficulties arose in cross-validating and generalizing instruments to new settings and different populations. More fundamentally, the relative inability of the method to advance psychological theory was a severe disappointment. With the advent of construct validity (Cronbach & Meehl, 1955), it became difficult to advocate exclusive reliance on pure "blind empiricism" in test construction. Yet, empirical approaches are still in use; in fact, 17% of the *Psychological Assessment* sample relied primarily on criterion groups for item selection.

Certainly, it is important not to throw the baby out with the bathwater. Correlations of a test with theoretically relevant criteria still constitute crucial evidence of validity, and there is no reason to avoid examining these correlations even in the early stages of scale development. One very strong approach would be to administer the initial item pool to a large heterogeneous sample (e.g., one encompassing both normal range and clinical levels of the target construct). Then, one basis (among several) for selecting items would be the power of the items to differentiate appropriately between subgroups in the sample (e.g., normal vs. clinical, or between individuals with different behavioral patterns or diagnoses within the clinical range).

*Internal consistency methods.* Currently, the single most widely used method for item selection in scale development is some form of internal consistency analysis. For example, 32% of the *Psychological Assessment* sample used factor analysis, and an additional 17% used another variant of the internal consistency method. These non–factor-analytic analyses typically

used corrected item–total correlations to eliminate items that did not correlate strongly with the assessed construct. Appropriately, factor analytic methods were used most frequently when the target construct was conceptualized as multidimensional and, therefore, subscales were desired. Indeed, whenever factor analysis was used, the resulting instrument had subscales, although subscales sometimes were developed without benefit of factor analysis, usually through some combination of rational and internal consistency analyses. Because Floyd and Widaman's (1995) article in this special issue examines the role of factor analysis in scale development in detail, we focus here on only a few basic issues.

First, put simply, factor analytic results provide information, not answers or solutions. That is, factor analysis is a tool that can be used wisely or foolishly. Naturally, the better one understands the tool the more likely it is to be used wisely, so we strongly recommend that scale developers either educate themselves about the technique or consult with a psychometrician at each stage of the development process. The power of the technique is such that blind adherence to a few simple rules is not likely to result in a terrible scale, but neither is it likely to be optimal.

Second, there is no substitute for good theory and careful thought when using these techniques. To a considerable extent, internal consistency is always had at the expense of breadth, so simply retaining the 10 or 20 "top" items may not yield the scale that best represents the target construct. That is, the few items correlating most strongly with the assessed or (in the case of factor analysis) latent construct may be highly redundant with one another; consequently, including them all will increase internal consistency estimates but also will create an overly narrow scale that likely will not assess the construct optimally. We consider this "attenuation paradox" (Loevinger, 1954) in more detail later.

Similarly, if items that reflect the theoretical core of the construct do not correlate strongly with it in preliminary analyses, it is not wise simply to eliminate them without consideration of why they did not behave as expected. Other explanations (e.g., Is the theory inadequate? Is the item poorly worded? Is the sample nonrepresentative in some important way? Is the item's base rate too extreme? Are there too few items representing the core construct?) should be considered before such items are eliminated.

*Item response theory (IRT).* Although IRT is by no means new, it has only recently begun to capture general attention. IRT is based on the assumption that test responses reflect an underlying trait (or set of traits, although most users assume that a single dominant trait can explain most of the response variance) and, moreover, that the relation between response and trait can be described for each test item by a monotonically increasing function called an *item characteristic curve* (ICC). Individuals with higher levels of the trait have higher expected probabilities for answering an item correctly (in the case of an ability) or in the keyed direction (for traits related to personality or psychopathology), and the ICC provides the precise value of these probabilities for each level of the trait.

Once the item parameters have been established (actually, estimated) by testing on a suitably large and heterogeneous group, IRT methods offer several advantages to scale developers. First, the methods provide a statistic indicating the precision with which an individual respondent's trait level is estimated. Thus, for example, the user can know whether the scale provides more precise estimates of the trait at the lower, middle, or upper end of the distribution. Second, trait-level estimates can be made independent of the particular set of items administered, thus providing greater flexibility and efficiency of assessment than is afforded by tests in which the ICCs are unknown. This property permits the development of computer-adaptive tests, in which assessment is focused primarily on those items for which maximum discriminative ability lies close to the respondent's trait level.

Standard intelligence tests make use of this IRT feature in a basic way. That is, older individuals are not administered the first, very easy items for each subtest unless they fail on the first few items tested. Rather, it is assumed that they would pass these items and they are given credit for them. Similarly, when examinees fail a sufficient number of items on a subtest, they are not administered the remaining, more difficult items under the assumption that they would fail them also. Scales developed using IRT simply apply these same features in a more comprehensive and precise manner. Interested readers are referred to Hambleton, Swaminathan, and Rogers (1991) for a relatively nontechnical presentation of IRT principles and applications and to King, King, Fairbank, Schlenger, and Surface (1993), Reise and Waller (1993), and Reise, Widaman, and Pugh (1993) for recent discussions.

### *Initial Data Collection*

*Inclusion of comparison (anchor) scales.* In the initial round of data collection, it is common practice to administer the preliminary item pool without any additional items or scales. This practice is regrettable, however, because it does not permit examination of the boundaries of the target construct; as we discussed earlier, exploring these boundaries is absolutely critical to understanding the construct from both theoretical and empirical viewpoints. Just as the literature was reviewed initially to discover existing scales and concepts to which the target is expected to be related and from which it must be differentiated, marker scales assessing these other constructs should be included in the initial data collection. Too often test developers discover late in the process that their new scale correlates .85 with an existing measure.

*Sample considerations.* It can be very helpful to do some preliminary pilot-testing on moderately sized samples of convenience (e.g., 100–200 college students for testing item formats) before launching a major scale development project. However, it is likely that some basic item content decisions will be made after the first full round of data collection, decisions that will shape the future empirical and conceptual development of the scale. Therefore, after initial pilot-testing, it is very important to use a large and appropriately heterogeneous sample for the first major stage of scale development. On the basis of existing evidence regarding the stability and replicability of structural analyses (Guadagnoli & Velicer, 1988), we recommend that a minimum of 300 respondents be assessed at this stage. Moreover, if the scale is to be used in a clinical setting it is critical to obtain data on patient samples early on, rather than

rely solely on college students until relatively late in the development process. One reason for obtaining data on patient samples early on is because the target construct may have rather different properties in different samples. If this fact is not discovered until late in the development process, the utility of the scale may be seriously compromised.

### *Psychometric Evaluation*

*Analysis of item distributions.* Before conducting more complex structural analyses, scale developers should examine the response distributions of the individual items. In inspecting these distributions, two considerations are paramount. First, it is important to identify and eliminate items that have highly skewed and unbalanced distributions. In a true–false format, these are items that virtually everyone (e.g., 95% or more) either endorses or denies; with a Likert rating format, these are items to which almost all respondents respond similarly (e.g., "slightly agree"). Highly unbalanced items are undesirable for several reasons. First, when most respondents answer similarly, items convey little information. Second, owing to their limited variability, these items are likely to correlate weakly with other items in the pool and therefore will fare poorly in subsequent structural analyses. Third, as noted earlier, items with extremely unbalanced distributions can produce highly unstable correlational results. Comrey (1988), for instance, pointed out that if one individual answers false to two items, whereas the remaining 199 all answer true, the items will correlate 1.0 with one another. With a more normal distribution, a high correlation would indicate that the items are redundant and that one of them probably should be eliminated. However, in this case, if that one individual changed just one of those responses to true, the 1.0 correlation would disappear. Clearly, the normal decision-making rules cannot be applied in this situation.

However, before excluding an item on the basis of an unbalanced distribution, it is essential to examine data from diverse samples representing the entire range of the scale's target population. Most notably, many items will show very different response distributions across clinical and nonclinical samples. For instance, the item "I have things in my possession that I can't explain how I got" likely would be endorsed by very few undergraduates and, therefore, would show a markedly unbalanced distribution in a student sample. In an appropriate patient sample, however, this item may have a much higher endorsement rate and, in fact, may be useful in assessing clinically significant levels of dissociative pathology. Thus, it may be desirable to retain items that assess important construct-relevant information in one type of sample, even if they have extremely unbalanced distributions (and relatively poor psychometric properties) in others.

This brings us to the second consideration, namely, that it is desirable to retain items showing a broad range of distributions. In the case of true–false items, this means keeping items with widely varying endorsement percentages. The reason for this is that most constructs are conceived to be—and, in fact, are empirically shown to be—continuously distributed dimensions, and scores can occur anywhere along the entire dimension. Consequently, it is important to retain items that discriminate at different points along the continuum. For example, in assessing the broad personality dimension of extraversion, it clearly would be undesirable to retain only those items that discriminated extreme introverts from everyone else; rather, one should include at least some items that differentiate extreme introverts from mild introverts, mild introverts from mild extraverts, and mild extraverts from extreme extraverts. Similarly, returning to an earlier example, the item "I have things in my possession that I can't explain how I got" may be useful precisely because it serves to define the extreme upper end of the dissociative continuum (i.e., those who suffer from dissociative identity disorder).

This is, in fact, one of the key advantages offered by IRT (King et al., 1993; Reise & Waller, 1993; Reise et al., 1993). As noted earlier, IRT yields parameter estimates that specify the point in a continuum at which a given item is maximally informative. These estimates, then, can be used as a basis for choosing an efficient set of items that yield precise assessment across the entire range of the continuum. Naturally, this almost invariably leads to the retention of items with widely varying distributions.

*Unidimensionality, internal consistency, and coefficient alpha.* The next crucial stage is to conduct structural analyses to determine which items are to be eliminated from or retained in the item pool. This stage is most critical when the test developer is seeking to create a theoretically based measure of a target construct, so that the goal is to measure one thing (i.e., the target construct)—and only this thing—as precisely as possible. This goal may seem relatively straightforward, but it is readily apparent from the recent literature that it remains poorly understood by test developers and users. The most obvious problem is the widespread misapprehension that the attainment of this goal can be established simply by demonstrating that a scale shows an acceptable level of internal consistency reliability, as estimated by an index such as coefficient alpha (Cronbach, 1951) or K-R 20 (Kuder & Richardson, 1937). A further complication is the fact that there are no longer any clear standards regarding what level of reliability is considered acceptable. For instance, although Nunnally (1978) recommended minimum standards of .80 and .90 for basic and applied research, respectively, it is not uncommon for contemporary researchers to characterize reliabilities in the .60s and .70s as good or adequate (e.g., Dekovic, Janssens, & Gerris, 1991; Holden, Fekken, & Cotton, 1991).

More fundamentally, psychometricians long have disavowed the practice of using reliability indices to establish the homogeneity of a scale (see Boyle, 1991; Cortina, 1993; S. B. Green, Lissitz, & Mulaik, 1977). To understand why this is so, it is necessary to distinguish between internal consistency on the one hand and homogeneity or unidimensionality on the other. *Internal consistency* refers to the overall degree to which the items that make up a scale are intercorrelated, whereas *homogeneity* and *unidimensionality* indicate whether the scale items assess a single underlying factor or construct (Briggs & Cheek, 1986; Cortina, 1993; S. B. Green et al., 1977). As such, internal consistency is a necessary but not sufficient condition for homogeneity or unidimensionality. In other words, a scale cannot be homogeneous unless all of its items are interrelated, but as we illustrate later, a scale can contain many interrelated items and still not be unidimensional. Because theory-driven assess-

ment seeks to measure a single construct systematically, the test developer ultimately is pursuing the goal of homogeneity or unidimensionality rather than internal consistency per se.

Unfortunately, K-R 20 and coefficient alpha are measures of internal consistency rather than homogeneity and so are of limited utility in establishing the unidimensionality of a scale. Furthermore, they are ambiguous and imperfect indicators of internal consistency because they essentially are a function of two parameters: the number of test items and the average intercorrelation among the items (Cortina, 1993; Cronbach, 1951). That is, one can achieve a high internal consistency reliability estimate by having either many items or highly intercorrelated items (or some combination of the two). Whereas the degree of item intercorrelation is a straightforward indicator of internal consistency, the number of items is entirely irrelevant. In practical terms, this means that as the number of items becomes quite large, it is exceedingly difficult to avoid achieving a high reliability estimate. Cortina (1993), in fact, suggested that coefficient alpha is virtually useless as an index of internal consistency for scales containing 40 or more items.

Accordingly, the average interitem correlation (which is a straightforward measure of internal consistency) is a much more useful index than coefficient alpha per se (which is not). Thus, test developers should work toward a target mean interitem correlation rather than try to achieve a particular level of alpha. As a more specific guideline, we recommend that the average interitem correlation fall in the range of .15–.50 (see Briggs & Cheek, 1986). This rather wide range is suggested because the optimal value necessarily will vary with the generality versus specificity of the target construct. If one is measuring a broad higher order construct such as extraversion, a mean correlation as low as .15–.20 probably is desirable; by contrast, for a valid measure of a narrower construct such as talkativeness, a much higher mean intercorrelation (perhaps in the .40–.50 range) is needed.

As suggested earlier, however, the average interitem correlation alone cannot establish the unidimensionality of a scale; in fact, a multidimensional scale actually can have an acceptable level of internal consistency. Cortina (1993, Table 2), for instance, reported the example of an artificially constructed 18-item scale composed of two distinct 9-item groups. The items that made up each cluster were highly homogeneous and in each case had an average interitem correlation of .50. However, the two groups were made to be orthogonal, such that items in different clusters correlated zero with one another. Obviously, the scale was not unidimensional, but instead reflected two distinct dimensions; nevertheless, it had a coefficient alpha of .85 and a moderate mean interitem correlation of approximately .24.

This example clearly illustrates that one can achieve a seemingly satisfactory mean interitem correlation by averaging many high coefficients with many low ones. Thus, unidimensionality cannot be ensured simply by focusing on the mean interitem correlation; rather, it is necessary to examine the range and distribution of these correlations as well. Consequently, we must amend our earlier guideline to state that virtually all of the individual interitem correlations should fall somewhere in the range of .15 to .50. Put another way, to ensure unidimensionality, almost all of the interitem correlations should be moderate in magnitude and should cluster narrowly around the mean value. B. F. Green (1978) articulated this principle most eloquently, stating that the item intercorrelation matrix should appear as "a calm but insistent sea of small, highly similar correlations" (pp. 665–666).

*The "attenuation paradox."* Some readers may be puzzled by our assertion that all of the interitem correlations should be moderate in magnitude. As we have seen, estimates of internal consistency will increase as the average interitem correlation increases; obviously, therefore, one can maximize internal consistency estimates by retaining items that are very highly correlated with others in the pool. It is not desirable, therefore, to retain highly intercorrelated items in the final scale?

No, it is not. This is the essence of the classic attenuation paradox in psychometric theory (see Boyle, 1991; Briggs & Cheek, 1986; Loevinger, 1954, 1957). Simply put, the paradox is that increasing the internal consistency of a test beyond a certain point will not enhance its construct validity and, in fact, may occur at the expense of validity. One reason for this is that strongly intercorrelated items are highly redundant: Once one of them is included in the scale, the other(s) contribute virtually no incremental information. For instance, it is well known that a test developer can achieve a highly reliable scale simply by writing several slightly reworded versions of the same basic item. Consider, for example, the three items "I often feel uncomfortable at parties," "Large social gatherings make me uneasy," and "I usually feel anxious at big social events." Because virtually everyone will respond to these variants in the same way (e.g., they either will endorse or deny them all), the items together will yield little more construct-relevant information than any one item individually. Accordingly, a scale will yield far more information—and, hence, be a more valid measure of a construct—if it contains more differentiated items that are only moderately intercorrelated.

Note, moreover, that maximizing internal consistency almost invariably produces a scale that is quite narrow in content; if the scale is narrower than the target construct, its validity is compromised. For instance, imagine two investigators each developing measures of general negative affect. The first chooses terms reflecting a wide array of negative mood states (scared, angry, guilty, sad, and scornful), whereas the second selects various indicators of fear and anxiety (scared, fearful, anxious, worried, and nervous). The latter scale will yield a higher reliability estimate, in that it consists of more semantically similar (and, therefore, more strongly intercorrelated) items; clearly, however, the former scale is a more valid measure of the broad construct of general negative affect.

In light of this paradox, it becomes clear that the goal of scale construction is to maximize validity rather than reliability. This is not to say that internal consistency estimates are useless or inappropriate. Indeed, coefficient alpha and other indices of internal consistency convey very important information regarding the proportion of error variance contained in the scale (see Cortina, 1993), and it is always desirable to demonstrate that a scale possesses an adequate level of reliability. Following the general guidelines of Nunnally (1978), we recommend that scale developers always strive for a coefficient alpha of at least .80; if a new scale or subscale falls below this mark, then revision should be undertaken to try to raise reliability to an acceptable level. This may involve writing additional items for a too-brief

scale or eliminating weaker items from a longer one. Nevertheless, an overconcern with internal consistency per se can be counterproductive: Once this benchmark of .80 has been secured with an appropriate number of items (as low as 4 or 5 items for very narrow constructs up to about 35 items for broad dimensions), there is no need to strive for any substantial increases in reliability.

*Structural analyses in scale construction.* Given that internal consistency estimates are untrustworthy guides, how can one achieve the desired goal of a undimensional scale? How does one produce a"calm sea of highly similar correlations"? It is conceivable that this could be accomplished through a careful inspection of the item intercorrelation matrix, perhaps in conjunction with a standard reliability program (such as those contained in SAS and SPSS). However, as the pool of candidate items increases, this process becomes unwieldy. Note, for instance, that a pool of only 30 items generates 435 individual intercorrelations to be inspected and evaluated, and that a pool of 40 items produces nearly 800 item intercorrelations.

Consequently, psychometricians strongly recommend that the test developer begin by factor-analyzing the items (Briggs & Cheek, 1986; Comrey, 1988; Cortina, 1993; Floyd & Widaman, 1995). Unfortunately, many test developers are hesitant to use factor analysis, either because it requires a relatively large number of respondents or because it involves several perplexing decisions. Both these concerns are unwarranted. First, it is true that factor analysis requires a minimum of 200–300 respondents (Comrey, 1988; Guadagnoli & Velicer, 1988), but this ultimately is no more than is needed for any good correlational or reliability analysis. Second, although the factor analyst must make a number of tactical decisions (e.g., methods of factor extraction and rotation), these decisions typically have much less effect on the resulting factor structures than is commonly believed; in fact, factor structures have been shown to be highly robust across different methods of factor extraction and rotation (see Guadagnoli & Velicer, 1988; Snook & Gorsuch, 1989; Watson et al., 1994). Hence, there is no reason to avoid using factor techniques in the initial stages of item selection. Nevertheless, as we stated earlier, the more one knows about this technique, the greater the probability that it will be used wisely; therefore, it is important that test developers either learn about the technique or consult with a psychometrician during the scale development process.

A thorough discussion of factor analysis is beyond the scope of this article (see especially Floyd & Widaman, 1995), but we will offer a very brief sketch of how it can be used in item selection. For the sake of simplicity, we consider the case of constructing a single unidimensional measure. First, subject the items to either a principal factor analysis (strongly preferred by Comrey, 1988) or a principal components analysis (recommended by Cortina, 1993) and extract the first few factors (say, four or five); in this simplified case, there is no need to be concerned with rotation. Next, examine the loadings of items on the first unrotated factor or component, which can be viewed as a direct measure of the common construct defined by the item pool. Items that load weakly on this first factor (below .35 in a principal factor analysis or below .40 in a principal components analysis) tend to be modestly correlated with the others and are leading candidates for removal from the scale. Similarly, items that have stronger loadings on later factors also are likely candidates for deletion. Conversely, items that load relatively strongly on the first factor and relatively weakly on subsequent factors are excellent candidates for retention. Thus, factor analysis quickly enables one to generate testable hypotheses regarding which items are good indicators of the construct and which are not. These predictions then can be evaluated in subsequent correlational and reliability analyses, which also can be used to identify pairs of redundant, highly correlated items.

A well-designed factor analysis also can play a crucial role in enhancing the discriminant validity of a new measure. For instance, we noted earlier that many new scales are not clearly differentiable from the broad trait of neuroticism (negative affectivity), thereby lacking discriminant validity. The easiest way to avoid creating yet another neuroticism measure is to subject the items of the provisional scale—together with a roughly equal number of neuroticism items—to a joint factor analysis. In this instance, one would extract two factors and rotate them to "simple structure" (e.g., using varimax or promax). Ideally, the target scale items (but often only a subset thereof) will load strongly on one factor, whereas the neuroticism items will load highly on the other. If not, then the new scale apparently is indistinguishable from neuroticism and the situation is likely to be hopeless. If so, then items that load strongly on the provisional scale factor—but quite weakly on the neuroticism factor—are excellent candidates for retention; conversely, items with relatively high loadings on the neuroticism factor have poor discriminant validity and probably should be dropped. This procedure can be followed for any construct that needs to be differentiated from the target scale, as long as marker items assessing the construct have been included in the initial data collection. At this stage of development, confirmatory factor analytic techniques also can be used to evaluate interrelations among scale items and their discriminant validity in comparison with related measures (see Floyd & Widaman, 1995, for an expanded discussion of the role of confirmatory factor analytic techniques in scale construction).

*Creating subscales.* We conclude this section with a brief consideration of subscales. In using the term *subscales,* we are referring to a situation in which a set of related measures are designed both to be assessed and analyzed separately and also to be combined into a single overall score. In other words, subscales are hypothesized to be specific manifestations of a more general construct. Defined in this way, subscales are a popular and important feature of test construction, as illustrated by the fact that approximately 70% of the *Psychological Assessment* sample included subscale development.

Creating valid subscales is an exceptionally tricky process, so much so that it is difficult to believe that it can be accomplished without some variant of factor analysis.[2] Indeed, the test constructor resembles the legendary hero Odysseus, who had to

[2] We acknowledge that this statement reflects a modern prejudice. Loevinger, Gleser, and DuBois (1953) developed a technique for "maximizing the discriminating power of a multiple-score test" (p. 309) that achieves the same end. This technique also has the practical advantage of treating items as all-or-none units, thereby paralleling the way they typically are used in scoring scales; by contrast, factor analysis apportions the item variance among the extracted factors, which necessitates decisions regarding factor-loading cutoffs to retain or eliminate items.

steer a narrow course between the twin terrors of Scylla and Charybdis. On the one hand, it makes no psychometric sense to combine unrelated items or subscales into a single overall score (although many scales developed by criterion keying do, in fact, show this property; see Carver, 1989). Accordingly, the scale developer must establish that all of the items—regardless of how they are placed in the various subscales—define a single general factor. If they do not, then the items need to be split off into separate, distinct scales. On the other hand, it also makes no psychometric sense to take a homogeneous pool of substantially intercorrelated items and arbitrarily divide it into separate subscales (e.g., on the basis of apparent differences in content). Accordingly, the scale developer must demonstrate that the intrasubscale item correlations (i.e., among the items that make up each subscale) are systematically higher than the intersubscale item correlations (i.e., between the items of different subscales). If this condition cannot be met, then the subscales should be abandoned in favor of a single overall score.

To illustrate the test developer's dilemma, consider the example of a test composed of two 10-item subscales. Let us further assume that the average intercorrelation of the items that make up Subscale A is .40, whereas that for Subscale B is .35. If, on the one hand, the average correlation between the A items and the B items is near zero—such that the two subscales also are essentially uncorrelated—then there is no justification for combining them into a single overall score; rather, they simply should be analyzed as two distinct constructs. On the other hand, if the average correlation between the A items and the B items is much above .30, there is no justification for dividing the items into two arbitrary subscales; instead, they simply should be summed into a single 20-item score. In this hypothetical case, the test developer's task is to have the mean correlation between the A items and B items be significantly greater than zero but substantially less than the average within-subscale values (say, .20). Without the assistance of a sophisticated structural technique such as factor analysis, this truly is a formidable task. Finally, we emphasize again that in making the decision of whether subscales are warranted, both theoretical and empirical considerations should be brought to bear, and data from diverse samples representing the entire range of the scale's target population should be considered.

## External Validity: The Ongoing Process

Just as graduation is properly called *commencement* to emphasize that it signals a beginning as well as an end, the process that we have described represents the initial rather than the final steps in scale development, refinement, and validation. However, the quality of the initial stages has clear ramifications for those stages that follow. For example, if the target concept is clearly conceptualized and delineated initially, then the resulting scale more likely will represent a novel contribution to the assessment armamentarium. If a widely relevant range of content is included in the original item pool, then the scale's range of clinical utility will be more clearly defined. Similarly, if the scale has been constructed with a focus on unidimensionality and not just internal consistency, then the scale will identify a more homogeneous clinical group, rather than a heterogeneous group requiring further demarcation. Finally, if issues of convergent and discriminant validity have been considered from the outset, then it will be far easier to delineate the construct boundaries precisely and to achieve the important goal of knowing exactly what the scale measures and what it does not.

Previously, Jackson (1970) has written extensively about the role of external validity in scale development. Moreover, in this issue, Smith and McCarthy (1995) describe the later refinement stages in some detail, so we conclude by noting simply that both the target of measurement and measurement of the target are important for optimal scale development. That is, later stages will proceed more smoothly if the earlier stages have been marked by both theoretical clarity (i.e., careful definition of the construct) and empirical precision (i.e., careful consideration of psychometric principles and procedures). Thus, we leave the aspiring scale developer well begun but far less than half done.

## References

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

Angleitner, A., & Wiggins, J. S. (1985). *Personality assessment via questionnaires.* New York: Springer-Verlag.

Bentler, P. M. (1969). Semantic space is (approximately) bipolar. *Journal of Psychology, 71,* 33–40.

Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences, 3,* 291–294.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality, 54,* 106–148.

Carver, C. S. (1989). How should multifaceted personality constructs be tested? Issues illustrated by self-monitoring, attributional style, and hardiness. *Journal of Personality and Social Psychology, 56,* 577–585.

Clark, L. A. (1993). *Schedule for Nonadaptive and Adaptive Personality (SNAP).* Minneapolis: University of Minnesota Press.

Clark, L. A., Watson, D., & Reynolds, S. (1995). Diagnosis and classification in psychopathology: Challenges to the current system and future directions. *Annual Review of Psychology, 46,* 121–153.

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology, 56,* 754–761.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98–104.

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

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

Dekovic, M., Janssens, J. M. A. M., & Gerris, J. R. M. (1991). Factor structure and construct validity of the Block Child Rearing Practices Report (CRPR). *Psychological Assessment, 3,* 182–187.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286–299.

Gangestad, S. W., & Snyder, M. (1991). Taxonomic analysis redux: Some statistical considerations for testing a latent class model. *Journal of Personality and Social Psychology and Social, 61,* 141–161.

Green, B. F., Jr. (1978). In defense of measurement. *American Psychologist, 33,* 664–670.

Green, D. P., Goldman, S. L., & Salovey, P. (1993). Measurement error

masks bipolarity in affect ratings. *Journal of Personality and Social Psychology, 64,* 1029–1041.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37,* 827–838.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Pscyhological Bulletin, 103,* 265–275.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment. A functional approach to concepts and methods. *Psychological Assessment, 7,* 238–247.

Hogan, R. T. (1983). A socioanalytic theory of personality. In M. Page (Ed.), *1982 Nebraska Symposium on Motivation* (pp. 55–89). Lincoln: University of Nebraska Press.

Holden, R. R., Fekken, G. C., & Cotton, D. H. G. (1991). Assessing psychopathology using structured test-item response latencies. *Psychological Assessment, 3,* 111–118.

Jackson, D. N. (1970). A sequential system for personality scale development. In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology* (Vol. 2, pp. 61–96). New York: Academic Press.

Jaffe, A., & Kilbey, M. M. (1994). The Cocaine Expectancy Questionnaire (CEQ): Construction and predictive utility. *Psychological Assessment, 6,* 18–26.

John, O. P. (1990). The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 66–100). New York: Guilford Press.

King, D. W., King, L. A., Fairbank, J. A., Schlenger, W. E., & Surface, C. R. (1993). Enhancing the precision of the Mississippi Scale for Combat-Related Posttraumatic Stress Disorder: An application of item response theory. *Psychological Assessment, 5,* 457–471.

Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design.* New York: Methuen.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151–160.

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

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635–694.

Loevinger, J., Gleser, G. C., & DuBois, P. H. (1953). Maximizing the discriminating power of a multiple-score test. *Psychometrika, 18,* 309–317.

Meehl, P. E. (1945). The dynamics of structured personality tests. *Journal of Clinical Psychology, 1,* 296–303.

Meehl, P. E., & Golden, R. R. (1982). Taxometric methods. In P. C. Kendall & J. N. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 127–181). New York: Wiley.

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

Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology, 65,* 143–151.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552–566.

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment, 7,* 300–308.

Snook, S. C., & Gorsuch, R. L. (1989). Component analysis versus common factor analysis: A Monte Carlo study. *Psychological Bulletin, 106,* 148–154.

Tellegen, A., & Waller, N. G. (in press). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In S. R. Briggs & J. M. Cheek (Eds.), *Personality measures: Development and evaluation* (Vol. 1). Greenwich, CT: JAI Press.

Turner, S., Beidel, D. C., Dancu, C. V., & Stanley, M. A. (1989). An empirically derived inventory to measure social fears and anxiety: The Social Phobia and Anxiety Inventory. *Psychological Assessment, 1,* 35–40.

Watson, D., & Clark, L. A. (1984). Negative affectivity: The disposition to experience aversive emotional states. *Psychological Bulletin, 96,* 465–490.

Watson, D., Clark, L. A., & Harkness, A. R. (1994). Structures of personality and their relevance to psychopathology. *Journal of Abnormal Psychology, 103,* 18–31.

Received March 30, 1995
Revision received April 4, 1995
Accepted April 4, 1995 ■