CHAPTER 14

The Construct Validation Approach to Personality Scale Construction

Leonard J. Simms
David Watson

Scale construction continues to be a popular activity among basic and applied personality researchers. We conducted a PsycINFO search of English-language journal articles published during the past 55 years that (1) included the keywords test construction, scale construction, or measure development and (2) also included the keyword personality. Using these criteria, our search revealed a total of 5,071 articles published since 1950, of which 3,609 (69.4%) have been published since 1985. Through the late 1980s and the 1990s, approximately 168 such articles, on average, were published each year, but this number has increased markedly in first half of this decade. Between the years 2000 and 2004, an average of 218 personality scale construction articles were published each year, representing a 30% increase as compared with the 15 years prior.

Several points are notable from these data. First, approximately two-thirds of all personality scale construction articles have been published over the past 20 years, likely reflecting both a resurgence of personality-based research and the proliferation of psychology journals in general. Second, although stable between 1985 and 1999, the pace of such publications appears to be increasing of late. Moreover, even the most recent articles have used a wide variety of approaches to construct and validate personality measures, with many reporting inadequate or outdated methodology, suggesting that the need for sound scale construction resources has never been greater (Clark & Watson, 1995; Watson, 2006). Thus, the primary goal of this chapter is to review basic principles of personality scale construction and describe an integrative method for constructing objective personality measures under the broad umbrella of construct validity.

The confusion often observed in the scale construction literature is not surprising when one considers the limited, and often outdated,
articles have been published, likely reflecting personality-based research and assessment texts. In most texts, methods of personality scale construction are described through a discussion of various specific scale construction approaches or strategies. In particular, many texts organize these strategies into those based on (1) rational or theoretical justifications, (2) empirical criterion-keying, and (3) factor analytic and internal consistency methods (e.g., Anastasi & Urbina, 1997; Kaplan & Saccuzzo, 2005)—which usually are described as mutually exclusive methods. As we discuss later, each approach carries clear strengths and limitations relative to the others. However, the combination of these approaches into a more integrative method of scale construction capitalizes on the unique strengths of each and makes it more likely that resultant measures will evidence adequate construct validity.

But what is "construct validity"? Often misunderstood and oversimplified, the concept of construct validity was first articulated in a seminal article by Cronbach and Meehl (1955), who argued that explicating the construct validity of a measure involves at least three steps: (1) describing a theoretical model—what Cronbach and Meehl called "the nomological net"—consisting of one or more hypothetical constructs and their relations to one another and to observable criteria, (2) building measures of the constructs identified by the theory, and (3) empirically testing the hypothesized relations between the constructs and observable criteria as specified by the theoretical model.

Different scale construction approaches tend to favor some aspects of the construct validation process while ignoring others. For example, measures derived using purely rational-theoretical methods may have direct connections to a clear, well-defined theory of a construct, but often fail to yield a clean pattern of convergent and discriminant relations when compared with other measures and with observable, non-test criteria. In contrast, the empirical criterion-keying approach results in measures that may reliably predict observable criteria but are devoid of any connection to theory.

How is construct validity involved in the scale construction process? All too often, researchers consider construct validity only in a post hoc fashion, as something that one establishes after the test has been constructed. However, construct validation is more appropriately considered a process, rather than an endpoint to which one aspires (Clark & Watson, 1995; Loewenstein, 1957; Messick, 1995). To maximize the practical utility and theoretical meaningfulness of a measure, the concepts of construct validity articulated by Cronbach and Meehl (1955) should be consulted at all stages of the scale construction process, including initial conceptualization of the construct(s) to be measured, development of an initial item pool, creation of provisional scales, cross-validation and finalization of scales, and validation against other test and nontest indicators of the construct(s).

Moreover, construct validity is not a static quality of a test that can be "established" in a definitive way with a single study or even a series of studies. Rather, the process of construct validation is dynamic. As Cronbach and Meehl (1955) describe, "In one sense, it is naive to inquire 'Is this test valid?' One does not validate a test, but only a principle for making inferences. If a test yields many different types of inferences, some of them can be valid and others invalid" (p. 297). Thus, as new scales begin to be examined against observable criteria, some aspects of the theory that guided its construction likely will be supported. However, other aspects of the theory may be refuted, and in such cases one must decide whether the fault lies with the test or the theory. This can be a tricky issue. Clearly, one cannot discard years of empirical work supporting a given theory because of a single study of a new measure. However, scales constructed rigorously, in accordance with the principles described in this chapter, have the potential to highlight problems with our understanding of theoretical constructs and lead to alternative hypotheses to be tested in future studies.

In addition to construct validity, researchers often speak of many other forms of validity—such as content validity, face validity, convergent validity, discriminant validity, concurrent validity, and predictive validity—that often are described as independent properties of a given measure. Recently, however, growing consensus has emerged that construct validity is best understood as a single overarching concept (American Psychological Association, 1999; Messick 1995; Watson, 2006). Indeed, as stated in the revised Standards for Educational and Psychological Testing (American Psychological Association, 1999), "Validity is a unitary concept. It is the degree to which all the ac-
cumulated evidence supports the intended interpretation of test scores for the proposed purpose" (p. 11). Thus, the concept of construct validity not only encompasses any form of validity that is relevant to the target construct, but also subsumes all of the major types of reliability. In sum, construct validity "has emerged as the central unifying concept in contemporary psychometrics" (Watson, 2006).

Loevinger (1957) was the first to systematically describe a theory-driven method of test construction firmly grounded in the concept of construct validity. In her monograph, Loevinger distinguished between three aspects of construct validity that she termed substantive validity, structural validity, and external validity. She argued that "these three aspects are mutually exclusive, exhaustive of the possible lines of evidence for construct validity, and mandatory" (pp. 653-654) and are "closely related to three stages in the test construction process: construction of the pool of items, analysis of the internal structure of the pool of items and consequent selection of items to form a scoring key, and correlation of test scores with criteria and other variables" (p. 654). Modern application of Loevinger's test construction principles has been described in detail elsewhere (e.g., Clark & Watson, 1995; Watson, 2006). In this chapter, our goals are to (1) summarize the basic features of substantive, structural, and external validity in the test construction process, (2) discuss a number of personality-relevant examples, and (3) propose ways to integrate principles of modern measurement theory (e.g., item response theory) in the development of construct valid personality scales.

To illustrate key aspects of the scale construction process, we draw on a number of relevant examples, including a personality measure currently being constructed by one of us (L. J. S.). This new measure, provisionally called the Evaluative Person Descriptors Questionnaire (EPDQ), was conceived and developed to provide an enhanced understanding of the Positive Valence and Negative Valence factors of the Big Seven model of personality (e.g., Beren-Martínez & Waller, 2002; Saucier, 1997; Tellegen & Waller, 1987; Waller, 1999). Briefly, the Big Seven model builds on the lexical tradition in personality research, which generally has suggested that five broad factors underlie much of the variation in human personality (i.e., the Big Five, or five-factor model of personality).

However, Tellegen and Waller (1987; Waller, 1999) argued that restrictions historically imposed on the dictionary descriptors used to identify the Big Five model ignored potentially important aspects of personality, such as stable individual differences in mood states and self-evaluation. Their less restrictive lexical studies resulted in seven broad factors: the familiar Big Five dimensions, plus two evaluative factors—Positive Valence (PV) and Negative Valence (NV)—reflecting extremely positive (e.g., describing oneself as exceptional, important, smart) and negative (e.g., describing oneself as evil, immoral, disgusting) self-evaluations, respectively. To date, only one measure of the Big Seven exists in the literature, the Inventory of Personal Characteristics #7 (IPC-7; Tellegen, Grove, & Waller, 1991), and this measure includes only global indices of PV and NV. Thus, the EPDQ is being developed to (1) provide an alternative measure of PV and NV to be used in structural personality studies, and (2) explore the lower-order facet structure of these dimensions.

The Substantive Validity Phase: Construct Conceptualization and Item Pool Development

A flowchart depicting the scale construction process appears in Figure 14.1. In it, we divide the process into three general phases, corresponding to the three aspects of construct validation originally articulated by Loevinger (1957) and reiterated by Clark and Watson (1995). The first phase—substantive validity—is centered on the tasks of construct conceptualization and development of the initial item pool.

Review of Literature

The substantive phase begins with a thorough review of the literature to discover all previous attempts to measure and conceptualize the construct(s) under investigation. This step is important for a number of reasons. First, if this review reveals that we already have good, psychometrically sound measures of the construct, then the scale developer must ask himself or herself whether a new measure is, in fact, necessary and, if so, why. With the proliferati...
In order to develop all previous scales, a thorough review of the literature is necessary. This review must include the construct validity, discriminant validity, and construct validation of each of the constructs. The questions asked are: Can we define the construct(s)? Is there evidence of construct validity? Is there evidence of discriminant validity? Is there evidence of construct validation? This step is critical, as it will help to identify potential constructs that are not currently being measured.

Once the construct(s) are defined, the next step is to develop an item pool. This can be done through the use of previously validated items or by generating new items. The initial item pool is then pilot tested and/or expert reviewed. Any problems that arise are then addressed by modifying or adding items to the pool. This process continues until a final pool of items is developed.

The developed item pool is then used to create the scale. This can be done through the use of psychiatric evaluation and scale creation. The resulting scale is then collected, and any problems that arise are addressed. This process continues until a final scale is developed.

The final scale is then evaluated for convergent and discriminant validity through the use of criterion-related validity studies. The results of these studies are then used to determine the final changes to the scale. The final scale is then written and reported.

FIGURE 14.1. Flowchart depicting the substantive, structural, and external validity phases of construct valid personality scale development.
tion for a new measure should be very carefully considered.

However, the existence of psychometrically sound measures of the construct does not necessarily preclude the development of a new instrument. Are the existing measures perhaps based on a very different definition of the construct? Are the existing measures perhaps too narrow or too broad in scope as compared with one's own conceptualization of the construct? Or are new measures perhaps needed to help advance theory or to cross-validate the findings achieved using the established measure of the construct? In the early stages of EPDQ development, the literature review revealed several important justifications for a new measure. First, as described above, the single available measure of PV and NV included only broad scales of these constructs, with too few items to identify meaningful lower-order facets. Second, factor analytic studies seeking to clarify personality structure require more than single exemplars of the constructs under investigation to yield theoretically meaningful solutions. Thus, despite the existence of the IPC-7 to tap PV and NV, the decision to develop the EPDQ appeared justified, and formal development of the measure was undertaken.

Construct Conceptualization

The second important function of a thorough literature review is to develop a clear conceptualization of the target construct. Although one often has a general sense of the construct before starting the project, the literature review likely will reveal alternative conceptualizations of the construct, related constructs that potentially are important, and potential pitfalls to consider in the scale development process. Clark and Watson (1995) recommend writing out a formal definition of the target construct in order to finalize one's model of the construct and clarify its breadth and scope. For the EPDQ, formal definitions were developed for PV and NV that included not only the broad aspects of extremely positive and negative self-evaluations, respectively, but also potential lower-order components of each identified in the literature. For example, the concept of PV was refined by Benet-Martinez and Waller (2002) to include a number of subcomponents, such as self-evaluations of distinction, intelligence, and self-worth. Therefore, the conceptualization of PV was expanded for the EPDQ to include these potentially important facets.

Development of the Initial Item Pool

Once the justification for the new measure has been established and the construct formally defined, it is time to create the initial pool of items from which provisional scales eventually will be drawn. This is a critical step in the scale construction process. As Clark and Watson (1995) described, "No existing data-analytic technique can remedy serious deficiencies in an item pool" (p. 311). Thus, great care must be taken to avoid problems that cannot be easily rectified later in the process. The primary consideration during this step is to generate items sampling all content that potentially is relevant to the target construct. Loewinger (1957) provided a particularly clear description of this principle, saying that "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. 639).

Thus, overinclusiveness should characterize the initial item pool in at least two ways. First, the pool should be broader and more comprehensive than one's theoretical model of the target construct. Second, the pool should include some items that may ultimately be shown to be tangential or perhaps even unrelated to the target construct. Overinclusiveness of the initial pool can be particularly important later in the scale construction process when one is trying to establish the conceptual and empirical boundaries of the target construct(s). As Clark and Watson (1995) put it, "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" (p. 311).

Central to substantive validity is the concept of content validity. Haynes, Richard, and Kubany (1995) defined content validity as "the degree to which elements of an assessment instrument are relevant to and representative of the target construct for a particular assessment purpose" (p. 238). Within this definition, relevancia refers to the appropriateness of a measure's items for the target construct. When applied to the scale construction process, this principle suggests that all items in the finished measure should fall within the boundaries of the target construct. Thus, although the principle of overinclusiveness suggests that some items be included in the initial item pool that fall outside the boundaries of the target con-
Initial Item Pool

A new measure has construct formally defined by the initial pool of items. Occasionally, both the scale-level and the target construct are defined as a representation of the concept. The principle of content validity suggests that final decisions regarding scale composition should take the relevance of items into account (Haynes et al., 1995; Watson, 2006).

A second important principle highlighted by Haynes and colleagues (1995) definition is the concept of representativeness, which refers to the degree to which the item pool adequately samples content from all important aspects of the target construct. Representativeness includes at least two important considerations. First, the item pool should contain items reflecting all concept areas relevant to the target construct. To ensure adequate coverage, many psychometricians recommend creating formal subscales to tap each important content area within a domain. In the development of the EPDQ, for example, an initial sample of 320 items was written to assess all areas of content deemed important to PV and NV, given the various empirical and theoretical considerations revealed by the literature review. More specifically, the pool contained “homogeneous item composites” (HCOs; Hogan, 1983; Hogan & Hogan, 1992), tapping a variety of relevant content highlighted by the literature review, including depravity, distinction, self-worth, perceived stupidity/intelligence, perceived attractiveness, and unconventionality/peculiarity (see, e.g., Benet-Martinez & Waller, 2002; Sauerte, 1997).

A second aspect of the representativeness principle is that the initial pool should include items reflecting all levels of the trait that need to be assessed. This principle is most commonly discussed with regard to ability tests, wherein a range of item difficulties are included so that the instrument can yield equally precise scores along the entire ability continuum. In personality measurement, this principle often is ignored for a variety of reasons. Items with extreme endorsement probabilities (e.g., items with which nearly all individuals will either agree or disagree) often are removed from consideration because they offer relatively little information relevant to most people's standing on the dimension, especially for traits with normal or nearly normal distributions in the general population. However, many personality measures are used across a diverse array of respondents—including college students, community-dwelling adults, psychiatric patients, and incarcerated individuals—who may differ substantially in their average trait levels. Thus, the item pool should reflect the entire range of trait levels along which reliable measurement is desired. Notably, psychometric methods based on classical test theory—which currently inform most personality scale construction projects—usually favor selection of items with moderate endorsement probabilities. However, as we will discuss in greater detail later, item response theory (IRT; see, e.g., Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991) offers valuable tools for quantifying the “trait level” of the items in the pool.

Haynes and colleagues (1995) recommend that the relevance and representativeness of the item pool be formally assessed during the scale construction process, rather than in a post hoc manner. A number of approaches can be adopted to assess content validity, but most involve some form of consultation with experts who have special knowledge of the target construct. For example, in the early stages of development of a new measure of posttraumatic symptoms, one of us (L. J. S.) and his colleagues are in the process of surveying practicing psychologists in order to gauge the relevance of a broad range of items. We expect that these expert ratings will highlight the full range of item content deemed relevant to the experience of trauma and will inform all later stages of item writing and scale development.

Writing Clear Items

Basic principles of item writing have been detailed elsewhere (e.g., Clark & Watson, 1995; Conrey, 1988). However, here we briefly discuss two broad aspects of item writing: item clarity and response format. Unclear items can lead to confusion among respondents, which ultimately results in less reliable and valid measurement. Thus, items should be written using simple and straightforward language that is appropriate for the reading level of the measure's target population. Likewise, it is best to avoid using slang and trendy or colloquial expressions that may quickly become obsolete, as they will limit the long-term usefulness of the measure. Similarly, one should avoid writing complex or convoluted items that are difficult to read and understand. For example, "double-barreled" items—such as the true-false item "I would like the work of a librarian because of my generally aloof nature"—should be avoided, because they confound two different characteristics (1) enjoyment of library work and (2) perceptions of aloofness or introversion. How are individuals to answer if they agree with one aspect of the item but not the others?
other? Such dilemmas infuse unneeded error into the measure and ultimately reduce reliability and validity.

The particular phrasing of items also can influence responses and should be considered carefully. For example, Clark and Watson (1995) suggested that writing items with stems such as "I worry about . . . " or "I am troubled by . . . " will build a substantial neuroticism/negative affectivity component into a scale. In addition, many writers (e.g., Anastasi & Urbina, 1997; Comrey, 1988; Kaplin & Saccuzzo, 2005) recommend writing a mix of positively and negatively keyed items to guard against response sets characterized by acquiescence (i.e., yea-saying) or denial (i.e., nay-saying). In practice, however, this can be quite difficult for some constructs, especially when the low end of the dimension is not well understood.

It also is important to phrase items so that all targeted respondents can provide a reasonably appropriate response (Comrey, 1988). For example, items such as "I get especially tired after playing basketball" or "My current romantic relationship is very good" assume contexts or situations that may not be relevant to all respondents. Rewriting the items to be more context-neutral—for example, "I get especially tired after I exercise" and "I've been generally happy with the quality of my romantic relationships"—increases the applicability of the resulting measure. A related aspect of this principle is that items should be phrased to maximize the likelihood that individuals will be willing to provide a forthright answer. As Comrey (1988) put it, "Do not exceed the willingness of the respondent to respond. Asking a subject a question that he or she does not wish to answer can result in several possible outcomes, most of them bad" (p. 757). However, when the nature of the target construct requires asking about sensitive topics, it is best to phrase such items using straightforward, matter-of-fact, and nonpejorative language.

Choice of Response Format

The two most common response formats used in personality measures are dichotomous (e.g., true-false or yes-no) and polytomous (e.g., Likert-type rating scales) (see Clark & Watson, 1995, for an analysis of alternative but less frequently used response formats, such as checklists, forced-choice items, and visual analog scales). Dichotomous and polytomous formats each come with certain strengths and limitations to be considered. Dichotomously scored items often are less reliable than their polytomous counterparts, and scales composed of such items generally must be longer in order to achieve comparable scale reliabilities (e.g., Conroy, 1988). Historically, many personality researchers adopted dichotomous formats for easier scoring and analyses. However, the power of modern computers and the extension of many psychometric models to polytomous formats have made these advantages less important. Nevertheless, all other things being equal, dichotomous items take less time to complete than polytomous items; thus, given limited time, a dichotomous item format may yield more information (Clark & Watson, 1995).

Polytomous item formats can vary considerably across measures. Two key decisions to make are (1) choosing the number of response options to offer and (2) deciding how to label these options. Opinions vary widely on the optimal number of response options to offer. Some argue that items with more response options yield more reliable scales (e.g., Comrey, 1988). However, there is little consensus on the "best" number of options to offer, as the answer likely depends on the fineness of discriminations that participants are able to make for a given construct (Kaplin & Saccuzzo, 2005). Clark and Watson (1995) add, "Increasing the number of alternatives actually may reduce reliability if respondents are unable to make the more subtle distinctions that are required" (p. 313). Opinions also differ on whether to offer an even or odd number of response options. An odd number of response options may entice some individuals to avoid giving careful consideration to some items by responding neutrally with the middle option. For that reason, some investigators prefer using an even number of options to force respondents to provide a nonneutral response.

Response options can be labeled using one of several anchoring schemes, including those based on agreement (e.g., strongly disagree to strongly agree), degree (e.g., very little to quite a bit), perceived similarity (e.g., uncharacteristic of me to characteristic of me), and frequency (e.g., never to always). Which anchoring scheme to use depends on the nature of the construct and the phrasing of items. In this regard, the phrasing of items must be compatible with the response format that has been chosen. For example, frequency modifiers may be quite
useful for items using agreement-based Likert scales, but will be quite confusing when used with a frequency-based Likert scale. Consider the item "I frequently drink to excess." As a true–false or agreement-based Likert item, the addition of "frequently" clarifies the meaning of the item and likely increases its ability to discriminate between individuals high and low on the trait in question. However, using the same item with a frequency-based Likert scale (e.g., 1 = never, 2 = infrequently, 3 = sometimes, 4 = often, 5 = almost always) is confusing to individuals because the frequency of the sample behavior is sampled twice.

Pilot Testing

Once the initial item pool and all other scale features (e.g., response formats, instructions) have been developed, pilot testing in a small sample of convenience (e.g., 100 undergraduates) and/or expert review of the stimuli can be quite helpful. Such procedures can help identify potential problems—such as confusing items or instructions, objectionable content, or the lack of items in an important content area—before a great deal of time and money are expended to collect the initial round of formal scale development data.

The Structural Validity Phase: Psychometric Evaluation of Items and Provisional Scale Development

Loevinger (1957) defined the structural component of construct validity as "the extent to which structural relations between test items parallel the structural relations of other manifestations of the trait being measured" (p. 661). In the context of personality scale development, this definition suggests that the structural relations between test and nontest manifestations of the target construct should be parallel to the extent possible—what Loevinger called "structural fidelity"—and, ideally, this structure should match that of the theoretical model underlying the construct. According to this principle, for example, the nature and magnitude of relations between behavioral manifestations of extraversion (e.g., sociability, talkativeness, gregariousness) should match the structural relations between comparable test items designed to tap these same aspects of the construct. Thus, the first step is to develop an item selection strategy that is most likely to yield a measure with structural fidelity.

Rational–Theoretical Item Selection

Historically, item selection strategies have taken a number of forms. The simplest of these to implement is the rational–theoretical approach. Using this approach, the scale developer simply writes items that appear consistent with his or her particular theoretical understanding of the target construct, assuming, of course, that this understanding is completely correct. The simplicity of this method is quite appealing, and some have argued that scales produced on solely rational grounds yield equivalent validity as compared with scales produced with more rigorous methods (e.g., Bürisch, 1984). However, such arguments fail to account for other potential pitfalls associated with this approach. For example, although the convergent validity of purely rational scales can be quite good, the discriminant validity of such scales often is poor. Moreover, assuming that one's theoretical model of the construct is entirely correct is unrealistic and likely will result in a suboptimal measure.

For these reasons, psychometricians argue against adopting a purely rational item selection strategy. However, some test developers have attempted to make the rational–theoretical approach more rigorous through additional procedures designed to guard against some of the problems described above. For example, having experts evaluate the relevance and representativeness of the items (i.e., content validity) can help identify problematic aspects of the item pool, so that changes can be made prior to finalizing the measure (Haynes et al., 1995). In another application, Harkness, McNulty, and Ben-Porath (1995) described the use of replicated rational selection (RRS) in the development of the PSY-5 scales of the second edition of the Minnesota Multiphasic Personality Inventory (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). RRS involves asking many trained raters—who are given a detailed definition of the target construct—to select items from a pool that most clearly tap the construct, given their interpretations of the definition and the items. Then, only items that achieve a high degree of consensus make the final cut. Such techniques are welcome advances over purely rational methods, but problems with discriminant validity often still emerge unless additional psychometric procedures are employed.
Criterion-Keyed Item Selection

Another historically popular item selection strategy is the empirical criterion-keying approach, which was used in the development of a number of widely used personality measures, most notably the MMPI-2 and the California Psychological Inventory (CPI; Gough, 1987). In this approach, items are selected for a scale based solely on their ability to discriminate between individuals from a “normal” group and those from a prespecified criterion group (i.e., those who exhibit the characteristic that the test developer wishes to measure). In the purest form of this approach, item content is irrelevant. Rather, responses to items are considered samples of verbal behavior, the meanings of which are to be determined empirically (Meehl, 1945). Thus, if one wishes to create a measure of extraversion, one simply identifies groups of extraverts and introverts, administers a range of items to each, and identifies items, regardless of content, that extraverts reliably endorse but introverts do not. The ease of this technique made it quite popular, and tests constructed using this approach often show reasonable validity.

However, empirically keyed measures have a number of problems that limit their usefulness in many settings. An important limitation is that empirically keyed measures are entirely atheoretical and fail to help advance psychological theory in a meaningful way (Loevinger, 1957). Furthermore, scales constructed using this approach often are highly heterogeneous, making the proper interpretation of scores quite difficult. For example, tables in the manuals for both the MMPI-2 (Butcher et al., 1989) and CPI (Gough, 1987) reveal a large number of internal consistency reliability estimates below .60, with some as low as .35, demonstrating a pronounced lack of internal coherence for many of the scales. Similarly problematic are the high correlations often observed among scales within empirically keyed measures, reflecting poor discriminant validity (e.g., Simms, Casillas, Clark, Watson, & Doebbeling, 2005). Thus, for these reasons, psychometricians recommend against adopting a purely empirical item selection strategy. However, some limitations of the empirical approach may reflect problems in the way the approach was implemented, rather than inherent deficiencies in the approach itself. Thus, combining this approach with other psychometric item selection procedures—such as those focusing on internal consistency and content validity considerations—offers a potentially powerful way to create measures with structural fidelity.

Internal Consistency Approaches to Item Selection

The internal consistency approach actually represents a variety of psychometric techniques drawing from classical reliability theory, factor analysis, and more modern techniques such as IRT. At the most general level, the goal of this approach is to identify relatively homogeneous scales that demonstrate good discriminant validity. This usually is accomplished with some variant of factor or component analysis, often combined with classical and modern psychometric approaches to hone the factor-based scales. In developing the EPDQ, for example, the initial pool of 120 items was administered to a large sample and then factor analyzed to determine the most viable factor structure underlying the item responses. Proportional scales were then created based on the factor analytic results as well as reliability considerations. The primary strength of this approach is that it usually results in homogeneous and differentiable dimensions. However, nothing in the statistical program helps to label the dimensions that emerge from the analyses. Therefore, it is important to note that the use of factor analysis does not obviate the need for sound theory in the scale construction process.

Data Collection

Once an item selection strategy has been developed, the first round of data collection can begin. Of course, the nature of this data collection will depend somewhat on the item selection strategy chosen. In a purely random-theoretical approach to scale construction, the scale developer might choose to collect expert ratings of the relevance and representativeness of each candidate item and then choose items based primarily on these ratings. If developing an empirically keyed measure, the developer likely would collect self-ratings on all candidate items from groups that differ on the target construct (e.g., those high and low in PV) and then choose the items that reliably discriminate between the groups.

Finally, in an internal consistency approach, the typical goal of data collection is to obtain
Psychometric Evaluation of Items

Because the internal consistency approach is the most common method used in contemporary scale construction (see Clark & Watson, 1995), in this section we focus on psychometric techniques from this tradition. However, a full review of internal consistency techniques is beyond the scope of this chapter. Thus, here we briefly summarize a number of important principles of factor analysis and reliability theory, as well as more modern approaches such as IRT, and provide references for more detailed discussions of these principles.

Factor Analysis

The basic goal of any exploratory factor analysis is to extract a manageable number of latent dimensions that explain the covariations among the larger set of manifest variables (see, e.g., Comrey, 1988; Fabrigar, Wegener, MacCallum, & Strahan, 1999; Floyd & Widaman, 1995; Preacher & MacCallum, 2003). As applied to the scale construction process, factor analysis involves reducing the matrix of interitem correlations to a set of factors or components that can be used to form provisional scales. Unfortunately, there is a daunting array of choices awaiting the prospective factor analyst—such as choice of rotation, method of factor extraction, the number of factors to extract, and whether to adopt an exploratory or confirmatory approach—and many avoid the technique altogether for this reason. However, with a little knowledge and guidance, factor analysis can be used wisely as a valuable tool in the scale construction process. Interested readers are referred to detailed discussions of factor analysis (see, e.g., Comrey, 1988; Floyd & Widaman, 1995), and Preacher and MacCallum (2003).

Regardless of the specifics of the analysis, exploratory factor analysis is extremely useful to the scale developer who wishes to create homogeneous scales (i.e., scales that measure one thing) that exhibit good discriminant validity. For demonstration purposes, abridged results from exploratory factor analyses of the initial pool of EPDQ items are presented in Table 14.1. In this particular analysis, all 120 items were included, and five oblique (i.e., correlated) factors were extracted. We should note here that there is no gold standard for deciding how many factors to extract in an exploratory analysis. Rather, a number of techniques—such as the scree test, parallel analyses of eigenvalues, and fit indices accompanying maximum likelihood extraction methods—provide some guidance as to a range of viable factor solutions, which should then be studied carefully (for discussions of the relative merits of these approaches, see Fabrigar et al., 1999; Floyd & Widaman, 1995; Preacher & MacCallum, 2003). Ultimately, however, the most important criterion for choosing a factor structure is the psychological and theoretical meaningfulness of the resultant factors. In this case, five factors—tentatively labeled Distraction, Worthlessness, NV/Evil Character, Oddity, and Perceived Stupidity—were extracted from the initial EPDQ data because (1) the five-factor solution was among those suggested by preliminary analyses and (2) this solution yielded the most compelling factors from a psychological standpoint.

In the abridged EPDQ output, six markers are presented for each factor in order to demonstrate a number of points (note that these are not simply the best six markers of each factor). The first point is that the goal of such an analysis is not necessarily to form scales using the top markers of each factor. Doing so might seem intuitively appealing, because using only the best markers will result in a highly reliable scale. However, high reliability often is gained
at the expense of construct validity. This phenomenon is known as the attenuation paradox (Loevinger, 1954, 1957), and it reminds us that the ultimate goal of scale construction is validity. Reliability of measurement certainly is important, but excessively high correlations within a scale will result in a very narrow scale that may show reduced connections with other test and nontest exemplars of the same construct. Thus, the goal of factor analysis in scale construction is to identify a range of items within each factor to serve as candidates for scale membership. Table 14.1 includes a number of candidate items for each EPDQ factor, some good and some bad.

Good candidate items are those that load at least moderately (at least 1.351; see Clark & Watson, 1995) on the primary factor and only minimally on other factors. Thus, of the 30 candidate items listed, only 18 meet this criterion, with the remaining items loading moderately on at least one other factor. Bad items, in contrast, are those that either load weakly on the hypothesized factor or cross-load on one or more factors. However, poorly performing items should be carefully examined before they are removed completely from consideration, especially when an item was predicted a priori to be a strong marker of a given factor. A number of considerations can influence the performance of an individual item. One's theory can be wrong; the item may be poorly worded or have extreme endorsement properties (i.e., nearly all or none of the participants endorsed the item); or perhaps sample-specific factors are to blame.

### TABLE 14.1. Abridged Factor Analytic Results Used to Construct the Evaluative Traits Questionnaire

<table>
<thead>
<tr>
<th>Line</th>
<th>Item</th>
<th>Abbreviated item text</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>52</td>
<td>People admire things I've done.</td>
<td>.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>83</td>
<td>I have many special aptitudes.</td>
<td>.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>69</td>
<td>I am the best at what I do.</td>
<td>.68</td>
<td>.60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>48</td>
<td>Others consider me valuable.</td>
<td>.64</td>
<td>-.29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>106</td>
<td>I receive many awards.</td>
<td>.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>66</td>
<td>I am needed and important.</td>
<td>.55</td>
<td>-.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>118</td>
<td>No one would care if I died.</td>
<td>.59</td>
<td></td>
<td>.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>28</td>
<td>I am an unimportant person.</td>
<td>.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>13</td>
<td>I would describe myself as stupid.</td>
<td>.55</td>
<td></td>
<td></td>
<td>.29</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>64</td>
<td>I'm relatively insignificant.</td>
<td>.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>113</td>
<td>I have little to offer the world.</td>
<td>.50</td>
<td>-.29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>11</td>
<td>I would describe myself as depressed.</td>
<td>.34</td>
<td></td>
<td>.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>84</td>
<td>I enjoy seeing others suffer.</td>
<td>.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>90</td>
<td>I engage in evil activities.</td>
<td>.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>41</td>
<td>I am evil.</td>
<td>.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>100</td>
<td>I lie, cheat, and steal.</td>
<td>.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>95</td>
<td>When I die, I'll go to a bad place.</td>
<td>.23</td>
<td>.36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>I am a good person.</td>
<td>.26</td>
<td>-.25</td>
<td>-.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>14</td>
<td>I am odd.</td>
<td>.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>88</td>
<td>My behavior is strange.</td>
<td>.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>9</td>
<td>Others describe me as unusual.</td>
<td>.73</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>29</td>
<td>I have unusual beliefs.</td>
<td>.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>93</td>
<td>I think differently from everybody.</td>
<td>.33</td>
<td></td>
<td>.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>98</td>
<td>I consider myself normal.</td>
<td>.29</td>
<td></td>
<td>.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>45</td>
<td>Most people are smarter than me.</td>
<td>.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>94</td>
<td>It's hard for me to learn new things.</td>
<td>.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>110</td>
<td>My IQ score would be low.</td>
<td>.22</td>
<td>.48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>80</td>
<td>I have very few talents.</td>
<td>.27</td>
<td>.41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>104</td>
<td>I have trouble solving problems.</td>
<td>.41</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>Others consider me foolish.</td>
<td>.25</td>
<td>.31</td>
<td>.32</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Loadings < 1.20 have been removed.
For example, Item 110 of the EPDQ (line 27 of Table 14.1: “If I took an IQ test, my score would be low”) loaded as expected on the Perceived Stupidity factor, but also loaded secondarily on the Worthlessness factor. Because of its face validity conjunction with the Perceived Stupidity factor, this item was tentatively retained in the item pool pending its performance in future rounds of data collection. However, if the same pattern emerges in future data, the item likely will be dropped. Another problematic item was Item 11 [line 12 of Table 14.1: “I would describe myself as depraved”], which loaded predictably but weakly on the N/Vil'li Character factor, but also cross-loaded (more strongly) on the Worthlessness factor. In this case, the item will be reworded in order to amplify the “depraved” aspect of the item and eliminate whatever nonspecific aspects contributed to its cross-loading on the Worthlessness factor.

Internal Consistency and Homogeneity

Once a reduced pool of candidate items has been identified through factor analysis, additional item-level analyses should be conducted to hone the scale(s). In the service of structural fidelity, the goal at this stage is to identify a set of items whose intercorrelations match the internal organization of the target construct (Watson, 2006). Thus, for personality constructs—which typically are hypothesized to be homogeneous and internally coherent—this principle suggests that items tapping personality constructs also should be homogeneous and internally coherent. The goal of most personality scales, then, is to measure a single construct as precisely as possible. Unfortunately, many scale developers and users confuse two related but differentiated aspects of internal coherence—(1) internal consistency, as measured by indices such as coefficient alpha (Cronbach, 1951), and (2) homogeneity or unidimensionality—often using the former to establish the latter. However, internal consistency is not the same as homogeneity (see, e.g., Clark & Watson, 1995; Schmitt, 1996). Whereas internal consistency indexes the overall degree of interrelation among a set of items, homogeneity (or unidimensionality) refers to the extent to which all of the items on a given scale tap a single factor. Thus, although internal consistency is a necessary condition for homogeneity, it clearly is not sufficient (Watson, 2006).

Internal consistency estimators such as coefficient alpha are functions of two parameters: (1) the average interitem correlation and (2) the number of items on the scale. Because such estimators confound internal coherence with scale length, scale developers often use a variety of alternative approaches—including examination of interitem correlations (Clark & Watson, 1992) and conducting confirmatory factor analyses to test the fit of a single-factor model (Schmitt, 1996)—to assess the homogeneity of an item pool. Here, we focus on interitem correlations. To establish homogeneity, one must examine both the mean and the distribution of the interitem correlations. The magnitude of the mean correlation generally should fall somewhere between .15 and .30. This range is wide to account for traits of varying bandwidths. That is, relatively narrow traits—such as those in the provisional Perceived Stupidity scale from the EPDQ—should yield higher average interitem correlations than broader traits such as those in the overall PV composite scale of the EPDQ (which is composed of a number of narrow but related facets, including reversed-keyed Perceived Stupidity). Interestingly, the provisional Perceived Stupidity and PV scales yielded average interitem correlations of .45 and .36, respectively, which was only somewhat consistent with expectations. The narrow trait indeed yielded a higher average interitem correlation than the broader trait, but the difference was not large, suggesting either that (1) the PV item pool is not sufficiently broad or (2) the theory underlying PV as a broad dimension of personality requires some modification.

The distribution of the interitem correlations also should be inspected to ensure that all clusters narrowly around the average, inasmuch as wide variation among the interitem correlations suggests a number of potential problems. Excessively high interitem correlations suggest unnecessary redundancy in the scale, which can be eliminated by dropping one item from each pair of highly correlated items. Moreover, significant variability in the interitem correlations may be due to multidimensionality within the scale, which must be explored.

Although coefficient alpha is not a perfect index of internal consistency, it continues to provide a reasonable estimate of one source of scale reliability. Thus, alpha should be computed and evaluated in the scale development process. However, given our earlier discussion
of the attenuation paradox, higher alphas are not necessarily better. Accordingly, some psychometricians recommend striving for an alpha of at least .80 and then stopping, as adding items for the sole purpose of increasing alpha beyond this point may result in a narrower scale with more limited validity (see, e.g., Clark & Watson, 1995). Additional aspects of scale reliability—such as test-retest reliability (see, e.g., Watson, 2006) and transient error (see, e.g., Schmidt, Le, & Iles, 2003)—also should be evaluated in this phase of scale construction to the extent that they are relevant to the structural fidelity of the new personality scale.

Item Response Theory

IRT refers to a range of modern psychometric models that describe the relations between item responses and the underlying latent trait they purport to measure. IRT can be an extremely useful adjunct to other scale development methods already discussed. Although originally developed and applied primarily in the ability testing domain, the use of IRT in the personality literature recently has become more common (e.g., Reise & Weller, 2003; Summ & Clark, 2003). Within the IRT literature, a variety of one-, two-, and three-parameter models have been proposed to explain both dichotomous and polytomous response data (for an accessible review of IRT, see Embretson & Reise, 2000, or Morizot, Ainsworth, & Reise, Chapter 24, this volume). Of these, a two-parameter model—with parameters for item difficulty and item discrimination—has been applied most consistently to personality data. Item difficulty (also known as "threshold" or "location") refers to the point along the trait continuum at which a given item has a 50% probability of being endorsed in the keyed direction. High difficulty values are associated with items that have low endorsement probabilities (i.e., that reflect higher levels of the trait). Discrimination reflects the degree of psychometric precision, or information, that an item provides at its difficulty level.

The concept of information is particularly useful in the scale development process. In contrast to classical test theory—in which a constant level of precision typically is assumed across the entire range of a measure—the IRT concept of information permits the scale developer to calculate conditional estimates of measurement precision and generate item and test information curves that more accurately reflect reliability of measurement across all levels of the underlying trait. In IRT, the standard error of measurement of a scale is equal to the inverse square root of the trait level, \( SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \)

where \( SE(\theta) \) and \( I(\theta) \) are the standard error of measurement and test information, respectively, evaluated at a given level of the underlying trait. If, therefore, one wishes only to discriminate between individuals who are moderate or high on this dimension—which likely would be the case in clinical settings—or if the goal is to measure the construct equally precisely across all levels of the trait—which would be desirable for computerized adaptive testing—then items would need to be added to the scale that provide more information at trait levels greater than 1.0 (i.e., items reflecting the same construct but with lower response base rates). If, however, one wishes only to discriminate between individuals who are low or moderate on the trait, then the current items may be adequate.

IRT also can be useful for examining the performance of individual items on a scale. Item information curves for five representative items...
accurately reflect across all levels of the standard error equal to the inverse square root of information at every point

standard error of measurement, regardless of the underlying standard errors of scores directly into or example, Fig. information and provisional Distinction this figure, the z-score metric, and the standard error of measurement is equal to the inverse square root of information at every point. The standard error axis is on the same metric as theta. This figure shows that measurement precision for this scale is greatest between theta values of -2.0 and +1.0.

FIGURE 14.2. Test information and standard error curves for the provisional EPDQ Distinction scale. Test information represents the sum of all item information curves, and standard error of measurement is equal to the inverse square root of information at all levels of theta. The standard error axis is on the same metric as theta. This figure shows that measurement precision for this scale is greatest between theta values of -2.0 and +1.0.

of the EPDQ Distinction scale are presented in Figure 14.3. These curves illustrate several notable points. First, not all items are created equal. Item 63 ("I would describe myself as a successful person"), for example, yielded excellent measurement precision along much of the trait dimension (range = -2.0 to +1.0), whereas Item 103 ("I think outside the box") produced an extremely flat information curve, suggesting that it is not a good marker of the underlying dimension. This is particularly interesting, given that the structural analyses that guided construction of this provisional scale identified Item 103 as a moderately strong marker of the Distinction factor. In light of these IRT analyses, this item likely will be removed from the provisional scale. Item 86 ("Among the people around me, I am one of the best"), however, also yielded a relatively flat information curve but provided incremental information at the very high end of the dimension. Therefore, this item was tentatively retained, pending the results from future data collection.

IRT methods also have been used to study item bias, or differential item functioning (DIF). Although DIF analyses originally were developed for ability testing applications, these methods have begun to appear more often in the personality testing literature to identify DIF related to gender (e.g., Smith & Reise, 1998), age cohort (e.g., Mackinnon et al., 1995), and culture (e.g., Huang, Church, & Katigbak, 1997). Briefly, the basic goal of DIF analyses is to identify items that yield significantly different difficulty or discrimination parameters across groups of interest, after equating the groups with respect to the trait being measured. Unfortunately, most such investigations are done in a post hoc fashion, after the measure has been finalized and published. Ideally, however, DIF analyses would be more useful during the structural phase of construct validation to identify and fix potentially problematic items before the scale is finalized.

A final application of IRT potentially relevant to personality is Computerized Adaptive Testing (CAT), in which items are individually tailored to the trait level of the respondent. A typical CAT selects and administers only those items that provide the most "psychometric in-
formation” at a given ability or trait level, eliminating the need to present items that have a very low or very high likelihood of being endorsed or answered correctly, given a particular respondent’s trait or ability level. For example, in a CAT version of a general arithmetic test, the computer would not administer easy items (e.g., simple addition) once it was clear from an individual’s responses that his or her ability level was far greater (e.g., he or she was correctly answering calculus or matrix algebra items). CAT methods have been shown to yield substantial time savings with little or no loss of reliability or validity in both the ability (Sand, Waters, & McBride, 1997) and personality (e.g., Simms & Clark, 2005) literatures.

For example, Simms and Clark (2005) developed a prototype CAT version of the Schedule for Nonadaptive and Adaptive Personality (SNAP; Clark, 1993) that yielded time savings of approximately 35% and 60%, as compared with full-scale versions of the SNAP completed via computer or paper-and-pencil, respectively. Interestingly, these data suggest that CAT, and nonadaptive computerized administration of questionnaires, offer potentially significant efficiency gains for personality researchers. Thus, CAT and computerization of measures may be attractive options for the personality scale developer that should be explored further.

The External Validity Phase: Validation against Test and Nontest Criteria

The final piece of scale development depicted in Figure 14.1 is the external validity phase, which is concerned with two basic aspects of construct validation: (1) convergent and discriminant validity and (2) criterion-related validity. Whereas the structural phase primarily involves analyses of the items within the new measure, the goal of the external phase is to examine whether the relations between the new measure and important test and nontest criteria are congruent with one’s theoretical understanding of the target construct and its place in the nomological net (Cronbach & Meehl, 1955). Data consistent with theory supports the construct validity of the new measure. However, discrepancies between observed data and theory suggest one of several conclusions—(1) the measure does not adequately
measure the target construct, (2) the theory requires modification, or (3) some of both—that must be addressed.

Convergent and Discriminant Validity

Convergent validity is the extent to which a measure correlates with other measures of the same construct, whereas discriminant validity is supported to the extent that a measure does not correlate with measures of other constructs that are theoretically or empirically distinct. Campbell and Fiske (1959) first described these aspects of construct validity and recommended that they be assessed using a multitrait-multimethod (MTMM) matrix. In such a matrix, multiple measures of at least two constructs are correlated and arranged to highlight several important aspects of convergent and discriminant validity.

A simple example—in which self-ratings and peer ratings of preliminary PV, NV, Extraversion, and Agreeableness scales are compared—is shown in Table 14.2. We must, however, exercise some caution in drawing strong inferences from these data, because the measures are not yet in their final forms. Nevertheless, these preliminary data help demonstrate several important aspects of an MTMM matrix.

First, the underlined values in the lower-left block are convergent validity coefficients comparing self-ratings on all four traits with their respective peer ratings. These should be positive and at least moderate in size. Campbell and Fiske (1959) summarized: "The entries in the validity diagonal should be significantly different from zero and sufficiently large to encourage further examination of validity" (p. 82). However, the absolute magnitude of convergent correlations will depend on specific aspects of the measures being correlated. For example, the concept of method variance suggests that self-ratings of the same construct generally will correlate more strongly than will self-ratings and peer ratings. In our example, the convergent correlations reflect different methods of assessing the constructs, which is a stronger test of convergent validity.

Ultimately, the power of an MTMM matrix lies in the comparisons of convergent correlations with other parts of the table. The ideal matrix would include convergent correlations that are greater than all other correlations in the table, thereby establishing discriminant validity, but three specific comparisons typically are made to explicate this issue more fully. First, each convergent correlation should be higher than other correlations in the same row and column in same box. Campbell and Fiske (1959) labeled the correlations above and below the convergent correlations heterotrait-heteromethod triangles, noting that convergent validity correlations "should be higher than the correlations obtained between that variable and any other variable having neither trait nor method in common" (p. 82). In Table 14.2, this rule was satisfied for Extraversion and, to a lesser extent, Agreeableness, but PV and NV clearly have failed this test of discriminant validity. The data are particularly striking for PV, revealing that peer ratings of PV actually correlate more strongly with self-ratings of NV and

<table>
<thead>
<tr>
<th>TABLE 14.2. Example of Multitrait-Multimethod Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Method</strong></td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>Self-ratings</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Peer ratings</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Note. N = 165. Correlations above .20 are significant at p < .01. Alpha coefficients are presented in parentheses along the diagonal. Convergent correlations are underlined. PV = positive valence; E = Extraversion; NV = negative valence; A = Agreeableness.
Agreeableness than with self-ratings of PV. Such findings highlight problems with either the scale itself or our theoretical understanding of the construct, which must be addressed before the scale is finalized.

Second, the convergent correlations generally should be higher than the correlations in the heterotrait-monomethod triangles that appear above and to the right of the heteromethod block just described. Campbell and Fiske (1959) described this principle by saying that a variable should "correlate higher with an independent effort to measure the same trait than with measures designed to get at different traits which happen to employ the same method" (p. 83). Again, the data presented in Table 14.2 provide a mixed picture with respect to this aspect of discriminant validity. In both the self-rating and peer-rating triangles, four of six correlations were significant and similar to or greater than the convergent validity correlations. In the self-rating triangle, PV and NV correlated -.38 with each other, PV correlated .48 with Extraversion, and NV correlated -.51 with Agreeableness, again suggesting poor discriminant validity for PV and NV. A similar, but more amplified, pattern emerged in the peer-rating triangle. Extraversion and Agreeableness, however, were uncorrelated with each other in both triangles, which is consistent with the theoretical assumption of the relative independence of these constructs.

Finally, Campbell and Fiske (1959) recommended that "the same pattern of trait interrelationship [should] be shown in all of the heterotrait triangles" (p. 83). The purpose of these comparisons is to determine whether the correlational pattern emerges regardless of method, then the former conclusion is plausible, whereas if significant differences emerge across the heteromethod triangles, then the influence of method variance must be evaluated. The four heterotrait triangles in Table 14.2 show a fairly similar pattern, with at least one key exception involving PV and Agreeableness. Whereas self-ratings of PV were uncorrelated with self-ratings and peer ratings of Agreeableness, peer ratings of PV strongly with self-ratings of Agreeableness. This finding is shown in Table 14.1, once sufficient data have been collected to construct validity of the provisional scales.

Criterion-Related Validity
A final source of validity evidence is criterion-related validity, which involves relating a measure to nontest variables deemed relevant to the target construct, given its nomological net. Most texts (e.g., Anastasi & Urbina, 1997; Kaplan & Saccuzzo, 2005) divide criterion-related validity into two subtypes based on the temporal relationship between the administration of the measure and the assessment of the criterion of interest. Concurrent validity involves relating a measure to criterion evidence collected at the same time as the measure itself, whereas predictive validity involves associations with criteria that are assessed at some point in the future. In either case, the primary goals of criterion-related validity are to (1) confirm the new measure's place in the nomological net and (2) provide an empirical basis for making inferences from test scores.

To that end, criterion-related validity evidence can take a number of forms. In the EPDQ development project, self-reported behavior data are being collected to clarify the behavioral correlates of PV and NV, as well as the facets of each. For example, to assess the concurrent validity of the provisional Perceived Stupidity facet scale, undergraduate participants in one study are being asked to report their current grade point averages. Pending these results, future studies may involve other related criteria, such as official grade point average data provided by the university, results from standardized achievement/aptitude test scores, or perhaps even individually administered intelligence test scores. Likewise, to examine the concurrent validity of the provisional Distinction facet scale, the same participants are being asked to report whether they have received any special honors, awards, or merit-based scholarships, or
should be finalized and the data published in a research article or test manual that thoroughly describes the methods used to construct the measure, appropriate administration and scoring procedures, and interpretable guidelines (American Psychological Association, 1999).

Summary and Conclusions

In this chapter we provide an overview of the personality scale development process in the context of construct validity (Cronbach & Meehl, 1955; Lovelingr, 1957). Construct validity is not a static quality of a measure that can be established in any definitive sense. Rather, construct validation is a dynamic process in which (1) theory and empirical work inform the scale development process at all phases and (2) data emerging from the new measure have the potential to modify our theoretical understanding of the target construct. Such an approach also can serve to integrate different conceptualizations of the same construct, especially to the extent that all possible manifestations of the target construct are sampled in the initial item pool. Indeed, this underscores the importance of conducting a thorough literature review prior to writing items and of creating an initial item pool that is strategically overinclusive. Lovelingr's (1957) classic three-part discussion of the construct validation process continues to serve as a solid foundation on which to build new personality measures, and modern psychometric approaches can be easily integrated into this framework.

For example, we discussed the use of IRT to help evaluate and select items in the structural phase of scale development. Although sparingly used in the personality literature until recently, IRT offers the personality scale developer a number of tools—such as detection of differential item functioning across groups, evaluation of measurement precision along the entire trait continuum, and administration of personality items through modern and efficient approaches such as CAT—which are becoming more accessible to the average psychometrician or personality scale developer. Indeed, most personality and social psychology journals now publish articles on IRT applications.

Recommended Readings


References


258 ASSESSING PERSONALITY AT DIFFERENT LEVELS OF ANALYSIS


