Social and Personality Psychology Compass 2/1 (2008): 414–433, 10.1111/j.1751-9004.2007.00044.x

Classical and Modern Methods of Psychological Scale Construction

Leonard J. Simms*
University at Buffalo, The State University of New York

Abstract

Scale construction is a growth enterprise in the psychological literature. Unfortunately, many measures promise much but are severely limited by the inadequacies of their conceptualization and execution. In this paper, a model for developing psychological scales is presented that is rooted in the traditions of construct validity and classical test theory but informed by modern psychometric methods. Construct validity is conceptualized as a guiding principle in each of three phases of scale development, focused on (i) construct conceptualization and development of the initial item pool, (ii) item selection and structural validity, and (iii) assessment of external validity vis-à-vis other measures and relevant nontest criteria.

Measurement of psychological constructs is an integral part of virtually all empirical and applied work in social and personality psychology, and in the field of psychology more generally. Such measures can take many different forms – such as psychophysiological apparatuses, structured interviews, collateral reports, behavioral observations, and implicit measures – but the most common form, and the focus of this paper, is the self-report method in which items are developed to tap psychological constructs of interest to us and then selected and grouped in some way to form scales. Development of such scales is quite popular, as evidenced by the increasing numbers of scale development papers that are published each year (Simms & Watson, 2007). The attraction to the self-report method is not difficult to understand. Such measures are relatively efficient compared to other methods, can be administered to large numbers of people with little cost, are easily scored, and often are the most direct methods for gathering information about people's thoughts, feelings, behavior, attitudes, and personality.

However, the apparent simplicity and efficiency of the method can be illusory, as much time, effort, and consideration are needed to develop measures that allow us to make reliable and valid inferences about people. Unfortunately, the literature is replete with measures that promise much but

provide little given the inadequacy of the psychometric methods used to develop and evaluate them (Clark & Watson, 1995; Simms & Watson, 2007; Watson, 2006). Thus, the primary aim of this paper is to summarize the currently accepted methods of psychological scale construction – under the broad umbrella of construct validity – and to discuss several modern psychometric methods that may be useful to those wishing to build or hone scales.

Traditional Models of Scale Construction

Methods of scale construction usually are organized into three mutually exclusive groups or strategies: (i) rational-theoretical approaches, (ii) empirical criterion keying, and (iii) factor-analytic and internal consistency methods. The rational-theoretical approach is the simplest method of scale construction, especially when it is used in it purest firm (i.e., without formal consideration of the psychometric properties of the scale). Using this approach, the scale developer simply writes items that appear consistent with his or her particular theoretical understanding of the target construct (i.e., items that have good face validity). The simplicity and efficiency of this method is quite attractive, with some arguing that rationally developed scales can yield equivalent validity compared to scales produced with more rigorous methods (e.g., Burisch, 1984). However, although convergent validity of purely rational scales can be quite good, this will not always be the case, and the discriminant validity of such scales often is poor. Moreover, rational-theoretical methods make the unrealistic assumption that one's theoretical model of the construct to be measured is wholly correct, which can result in measures with incomplete or inaccurate coverage of the construct. Thus, psychometricians usually argue against adopting a purely rational scale construction strategy.

The empirical criterion-keying method has been used to develop several widely used measures, such as the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1943; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) and the California Psychological Inventory (Gough, 1987). In this approach, items are selected for a scale based solely on their ability to discriminate between individuals from two groups of interest. For example, scale 2 of the MMPI was developed by contrasting the item responses of a 'normal' group with those from a criterion group of depressed patients; items that discriminated between the groups were considered for inclusion on the scale, without regard for item content. In this approach, responses to items are considered to be samples of verbal behavior, the meanings of which are to be determined empirically (Meehl, 1945). Thus, item content becomes much less relevant than in other methods of scale construction. Measures developed using criterion keying can show adequate convergent validity, but they often evidence a number of problems – such as poor internal coherence, poor discriminant validity, and the lack of theoretical roots or importance – that

limit their usefulness in many settings. Thus, most psychometricians also recommend against adopting a purely empirical scale construction strategy.

The third traditional method of scale construction is the internal consistency or factor-analytic approach. In this approach, the primary goal is to identify relatively homogenous scales that demonstrate good discriminant validity. This usually is accomplished with some variant of factor or component analysis, which is used to identify coherent dimensions among large numbers of items written to sample one or more candidate constructs to be measured. The primary strength of this approach is that it usually results in homogeneous and differentiable dimensions. However, nothing in the statistical program helps the user to label the dimensions that emerge from the analyses. As such, the use of factor analysis does not eliminate the need for sound theory in the scale construction process. Moreover, this approach assumes that the constructs we wish to measure are relatively homogenous in nature, which may not be true in all cases (e.g., constructs that are inherently multidimensional, like some psychiatric syndromes). Thus, although factor analysis is an important piece in most scale construction strategies, strict adherence to a pure internal consistency approach usually is not ideal.

Construct Validity as a Unifying Framework

Thus, each of the traditional scale construction approaches carries clear strengths and limitations relative to the others. As such, many psychometricians argue that an integrative approach is most optimal, one in which the relevant aspects of both classical and modern psychometric methods are combined in the service of building measures that maximize construct validity (e.g., Loevinger, 1957; Clark & Watson, 1995; Simms & Watson, 2007). In their seminal paper on construct validity, Cronbach and Meehl (1955) argued that establishing the validity of measures of psychological constructs is challenging because there are no clear, observable criteria to serve as gold standards for the constructs we wish to measure. As a result, the process of construct validation requires that measures of such constructs be embedded in a theoretical network of predicted relations among hypothetical constructs and observable criteria. Such a network then permits theory-driven investigations of a measure's reliability and validity. All too often, unfortunately, scale developers consider construct validity only *after* a scale has been constructed. However, to maximize the practical utility and theoretical meaningfulness of a measure, the concepts of construct validity articulated by Cronbach and Meehl are better considered at all stages of the scale construction process (Clark & Watson; Loevinger, 1957; Messick, 1995; Simms & Watson, 2007).

In a seminal paper, Loevinger (1957) described a theory-driven approach to scale construction in which she distinguished among three aspects of construct validity – *substantive validity*, *structural validity*, and

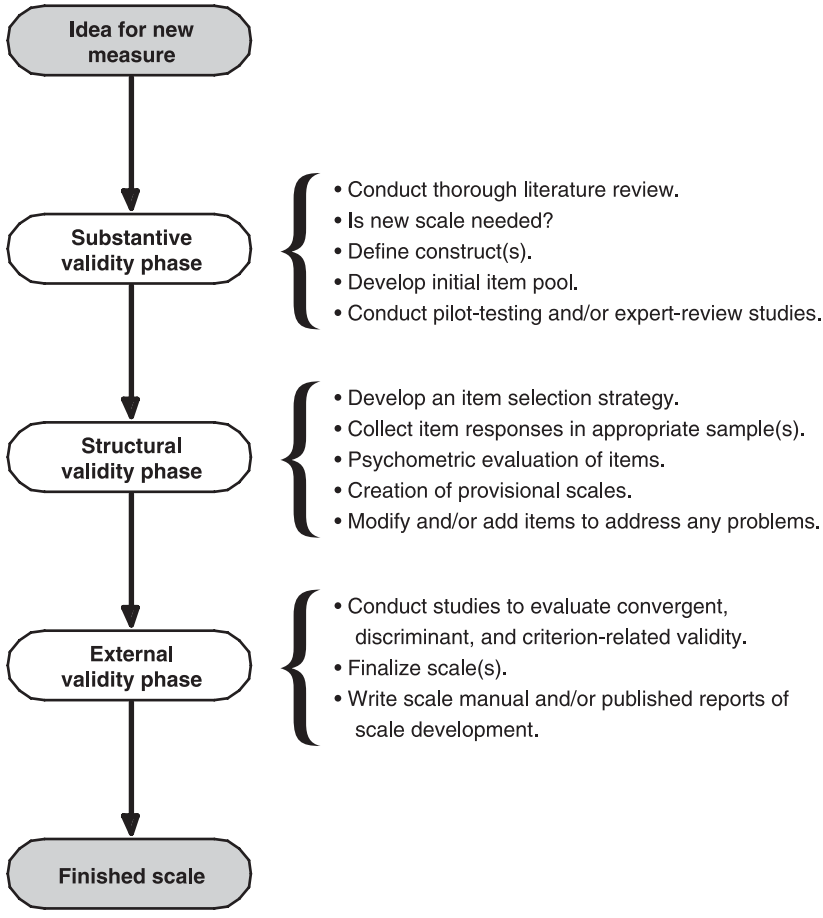

Figure 1 Flowchart depicting the phases of scale development.

external validity – that serve as a process through which construct valid scale development may take place. In this paper, the goals are to (i) summarize the basic features of each phase of Loevinger's model of scale construction, and (ii) discuss ways to integrate principles of modern psychometric theory (e.g., item response theory) with traditional methods based on classical test theory throughout the scale construction process.

Substantive Validity Phase

Figure 1 includes a brief flowchart depicting the scale construction process. Once one decides to develop a new measure, the first step is to complete the substantive validity phase, the primary goals of which are to form clear, theory-informed conceptualizations of all constructs to be measured and to develop the initial item pool. This process begins with a thorough

literature review to identify all previous attempts to measure and conceptualize the construct(s) under investigation. This step serves two important purposes. First, because the literature is full of measures for nearly all conceivable constructs, the review should investigate whether a new measure truly is needed. If, for example, the literature review reveals that other psychometrically sound measures already exist for the construct, then the prospective scale developer must then either identify ways in which the new measure will improve on previous attempts or drop the project altogether. The second important function of the literature review is to develop a clear conceptualization of the target construct(s) to be measured. The literature review usually will reveal alternative conceptualizations of the constructs, related constructs that potentially are important, and potential pitfalls to consider in the scale development process. After completing the review, formal definitions should be written for each target construct that serve to clarify the breadth and scope of each and inform item development (Clark & Watson, 1995; Haynes, Richard, & Kubany, 1995; Simms & Watson, 2007).

The next step is to develop an initial item pool. This is an important step, since serious problems with the item pool will reverberate through all subsequent data analyses and scale construction efforts. As Clark and Watson (1995, 311) noted, 'No existing data-analytic technique can remedy serious deficiencies in an item pool.' In short, the primary goal is to generate an item pool with good content validity (Haynes et al., 1995; Loevinger, 1957). That is, items should be written that are (i) *relevant* to the constructs to be measured, and (ii) *representative* of all potentially important aspects of the target construct. Having formal construct definitions is particularly important here, as such definitions should guide the item writing process. Loevinger (1957, 659) argued that the item pool should be purposely overinclusive at this stage, such 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.' Following this advice will help define the conceptual and empirical boundaries of the target constructs and increase the likelihood that all relevant content was included in the item pool.

In many scale construction projects, the representativeness principle encourages us to include items relevant to all possible facets of a given construct. For example, for a measure of a broad construct like extraversion, many would argue that items should be written to assess all elements of the construct, such as sociability, dominance, positive affect, exhibitionism, talkativeness, etc. However, a second aspect of the representativeness principle is that the item pool should include content reflecting all levels of the trait that need to be reliably assessed (Simms & Watson, 2007). This concept is most clear if one considers a traditional ability test in which, for example, the goal is to measure mathematical ability. On such a test, the nature of the items would depend on the range of the underlying

ability one wishes to capture reliably. If the goal is to measure mathematical ability in entering college students, then the item pool likely will need to include relatively difficult items (e.g., geometry, trigonometry, calculus), whereas a mathematics placement test for junior high school would require less difficult content (e.g., addition, multiplication, fractions, decimals).

Although often ignored in personality measurement, this principle is quite important and should be considered when developing the item pool. Many personality measures are used across a wide variety of individuals – including college students, community adults, psychiatric patients, and incarcerated individuals – who likely differ substantially in their average trait levels. As such, items should be included in the pool to reflect all different manifestations and levels of the underlying trait for which reliable measurement is desired. For example, if one wishes to measure aggression across a wide variety of settings, items such as 'I become angry more often than I like' might help discriminate between those low and moderate in trait aggression, whereas items such as 'I get into a lot of fistfights' likely will be more helpful in discriminating between those moderate and high on the trait. Unfortunately, classical scale construction methods generally favor items with moderate endorsement probabilities; thus, more severe items often are tossed out prematurely despite their possible relevance to the extreme ends of the trait dimension. However, modern psychometric models, such as those based in item response theory (IRT; e.g., Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991) can ameliorate this problem to a large extent and offer tools for quantifying the 'trait level' associated with items in the pool.

But how should one go about writing the initial set of items? A variety of sources have described the basic principles of item writing in much greater detail than can be done here (e.g., Anastasi & Urbina, 1997; Clark & Watson, 1995; Comrey, 1988; Kaplan & Saccuzzo, 2005; Simms & Watson, 2007). Most such principles boil down to two main topics: item clarity and response format. Unclear or grammatically problematic items can lead to comprehension problems among respondents that ultimately result in less reliable and valid measurement. Moreover, careful attention should be paid to the nature and number of response options provided to respondents, as such features will influence the ways items must be written. These basic item-writing and response option principles are elaborated with specific suggestions in Table 1.

Finally, when 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 a lack of items in an important content area – before extensive time and money are expended to collect the initial round of formal scale development data.

 Social and Personality Psychology Compass 2/1 (2008): 414–433, 10.1111/j.1751-9004.2007.00044.x

Table 1 Summary of basic item writing principles

No.	Item writing guideline
1	Write items using simple and straightforward language that is appropriate for the reading level of the measure's target population.
2	Avoid writing complex or convoluted items that are difficult to read and understand (e.g., double-barreled items such as 'My outgoing nature would make me a good salesperson', since they confound different characteristics – in this case, being outgoing and being a good salesperson – that may not covary in some individuals.
3	Avoid using slang and colloquial expressions that may quickly become obsolete.
4	Be careful that phrasing does not affect responses in unexpected ways (e.g., including 'worry' in an item nearly guarantees that the it will have a neuroticism component).
5	To the extent possible, write a mix of positively and negatively worded items to guard against response sets.
6	Phrase items generally enough that most or all targeted respondents can provide a reasonably appropriate response (e.g., write 'I get tired after I exercise' rather than 'I get tired after playing soccer').
7	To increase the likelihood of truthful responding, phrase items asking about sensitive issues using straightforward, matter-of-fact, and nonpejorative language.
8	Choose the item response format carefully. Dichotomous items (e.g., true-false or yes-no) take less time to complete but generally are less reliable than polytomous items (e.g., Likert-type rating scales).
9	For polytomous items, carefully consider the number of response options to offer. More is not necessarily better, as respondents may not be able to reliably distinguish between the adjacent anchors on a Likert scale that is too finely graded.
10	Consider whether to provide an odd or even number of response options. An odd number may entice some to hastily respond with the middle option. An even number of options forces respondents to provide a non-neutral response.
11	Consider the anchoring scheme for response options. These can be based on agreement (e.g., *strongly disagree* to *strongly agree*), frequency (e.g., *never* to *always*), degree (e.g., *very little* to *quite a bit*), and perceived similarity (e.g., *uncharacteristic of me* to *characteristic of me*).
12	Be sure that item phrasing is consistent with the response option anchoring scheme (e.g., the item 'I often get into fistfights' would work fine with an agreement-based anchoring scheme but would be confusing with a frequency-based scheme).

Structural Validity Phase

The primary goals of the structural validity phase are to collect responses to the initial set of items, generate and implement an item selection strategy, and construct provisional scales. As described above, each of these goals are to be focused on the overarching goal of creating a measure with good construct validity. According to Loevinger (1957, 661), construct

validity is enhanced in the structural phase to the extent that 'structural relations between test items parallel the structural relations of other manifestations of the trait being measured,' something she called 'structural fidelity.' As such, the item selection strategy should be designed to maximize structural fidelity. For example, this principle implies that if one wishes to select items for facet scales of the broad domain of extraversion (e.g., sociability, dominance, positive affect, exhibitionism, talkativeness), the structural relations among test items reflective of these facets should match the structural relations among comparable nontest, behavioral manifestations of these same aspects of the construct.

Given the prominent trait-dimensional perspective underlying much of personality psychology today, the structural validity phase usually includes an item selection strategy focused on creating relatively homogeneous scales (i.e., scales that measure one thing) that are reasonably distinct from one another (i.e., exhibit good discriminant validity). Thus, factor analyses and other classical and modern psychometric procedures usually are the primary methods used to ensure structural fidelity. It is these methods that are summarized in this paper. Of course, if the theoretical/empirical structure underlying a given construct is something other than a continuous dimension (e.g., type models), then wholly different item selection methods would be needed to ensure structural fidelity in the resulting measure.

Prior to selecting items, an initial round of data collection is needed to gather responses to the initial item pool. In an internal consistency approach, the goal of data collection is to obtain self-ratings for all candidate items in a large sample representative of the population(s) in which the measure ultimately will be used. Many researchers use college undergraduates for the initial data collection since such samples can be collected efficiently and with little cost. However, if one wishes to develop a measure that generalizes beyond students, inclusion of a broader range of participants is desirable, even early in the data collection process. For example, for prospective measures of personality pathology, strict reliance on an undergraduate sample would not be appropriate. Instead, responses also should be collected from psychiatric and criminal samples in which personality pathology is more prevalent. Several rounds of data collection may be necessary before provisional scales are ready for the next phase of scale development. Between each round, psychometric analyses are used to begin building provisional scales and to identify problems with the item pool (e.g., poorly functioning items, gaps in content).

Item selection using exploratory factor analysis

Exploratory factor analysis (EFA) is among the most widely used classical tools for creating internally consistent scales. The basic goal of EFA is to extract a manageable number of latent dimensions that explain the covariation among a larger set of manifest variables (e.g., Comrey, 1988;

Fabrigar, Wegener, MacCallum, & Strahan, 1999; Floyd & Widaman, 1995; Preacher & MacCallum, 2003). The use of EFA comes with a large number of decisions that must be made (e.g., number of factors to extract, orthogonal versus oblique rotation, estimation of principal components versus common factors), and a detailed accounting of this procedure is beyond the scope of this paper. Interested readers are referred to detailed discussions of EFA procedures by Fabrigar et al. (1999), Floyd and Widaman, and Preacher and MacCallum. When used to develop scales, regardless of the specific procedures implemented, EFA involves reducing the matrix of interitem correlations to a set of factors or components to be used as a basis for forming provisional scales.

The results of an EFA come in the form of a factor loading matrix that includes the loadings of all items on all factors extracted. Given this information, how is one to choose items for scales? Several strategies may be used. A simplistic approach would be to form scales by choosing all of the highest loading items, regardless of their loadings on other factors. Although its simplicity may be appealing, this approach has several clear problems. First, if the initial item pool was rather large – which will often be the case when being appropriately overinclusive – scales formed in this manner may be excessively large and contain pockets of redundancy caused by highly similar items. Although using only the best markers will result in a highly reliable scale, high reliability often is gained at the expense of construct validity (see discussion of the attenuation paradox in Loevinger, 1954, 1957, and Clark & Watson, 1995), since excessively high correlations within a scale may result in a highly narrow scale that may show reduced connections with relevant criteria. Also, because some items will have nontrivial cross-loadings on other factors, this approach may cause unintended scale overlap.

Thus, a better approach is to identify a range of good items within each factor to serve as candidates for scale membership. Good candidate items are those that load at least moderately (at least $|0.35|$; see Clark and Watson, 1995) on the primary factor and only minimally on other factors. Poor items, in contrast, are those that either load weakly on their hypothesized factors or cross-load on more than one factor. However, one must be careful not to prematurely drop poorly performing items, especially when such items were predicted a priori to be strong markers of a given factor. Sampling error, for example, may be the culprit, and some problematic items may work much better in a new round of data collection. In general, it is best to base provisional scales on the responses of multiple samples and to identify items that perform robustly.

Building measurement precision and homogeneity into a scale

In addition to EFA, other internal consistency methods should be used to evaluate the provisional scales as they are iteratively developed. The goal

here is to ensure that the new scales are sufficiently coherent. Unfortunately, many researchers confuse two distinct components of internal coherence – (a) *internal consistency reliability*, measured by indices such as Cronbach's coefficient alpha, and (b) *homogeneity* or *unidimensionality*. However, internal consistency is not the same as homogeneity (e.g., Clark & Watson, 1995; Cortina, 1993; John & Soto, 2007; Schmitt, 1996). Whereas internal consistency reliability is a function of the average degree of correlation among a set of items as well as the total number of items, homogeneity refers to the extent to which all of the items on a given scale tap a single dimension. Indeed, it is known that alpha provides accurate estimates of internal consistency only under conditions that are seldom met in the social personality literature. For example, multidimensionality in the item pool will cause alpha to be an underestimate of reliability, whereas other sources of measurement error (e.g., both random and systematic) will cause alpha to be an overestimate of the reliability of the measure. Thus, both provide important information about a scale, but they are not interchangeable.

Practically speaking, reliability is important because unreliability in a scale (i) attenuates its validity correlations with criteria to which it might be compared (e.g., John & Soto, 2007), and (ii) severely limits the degree of confidence we can have about the magnitude of a given person's score (Nunnally, 1978). In classical test theory (Gulliksen, 1950), the latter concern is best quantified using the standard error of measurement (SEM):

$$SEM_{test} = SD_{test}\sqrt{1 - reliability_{test}},$$

where SEM_{test} is the standard error of measurement of a given test, and SD_{test} is the standard deviation of the test. This places the *SEM* on the same metric as the measure's *SD*, which is useful for interpretation. Figure 2 shows the theoretical value of *SEM* as well as the 95% and 99% confidence limits as a function of scale reliability. These values, which were computed using a *Z*-score metric, are somewhat sobering. At an alpha of 0.80, the 95% and 99% confidence intervals around a given score would be 0.88 and 1.15 standard deviation units, respectively. To make the impact of these values more clear, consider an observed score of 70 on the *T*-score metric (scores scaled to a normative *M* of 50 and a *SD* of 10) that often is used in applied personality assessment. Given these parameters, there is a 95% probability that this person's 'true score' lies somewhere between 61.2 and 78.8, and a 99% probability that the true score is between 58.5 and 81.5. The wideness of these intervals especially lowers our confidence when trying to make interpretive inferences based on the observed score. On the MMPI-2, for example, a *T* score of 65 is considered to be a clinically significantly elevation. However, although our hypothetical person's observed score *appears* to be clinically elevated

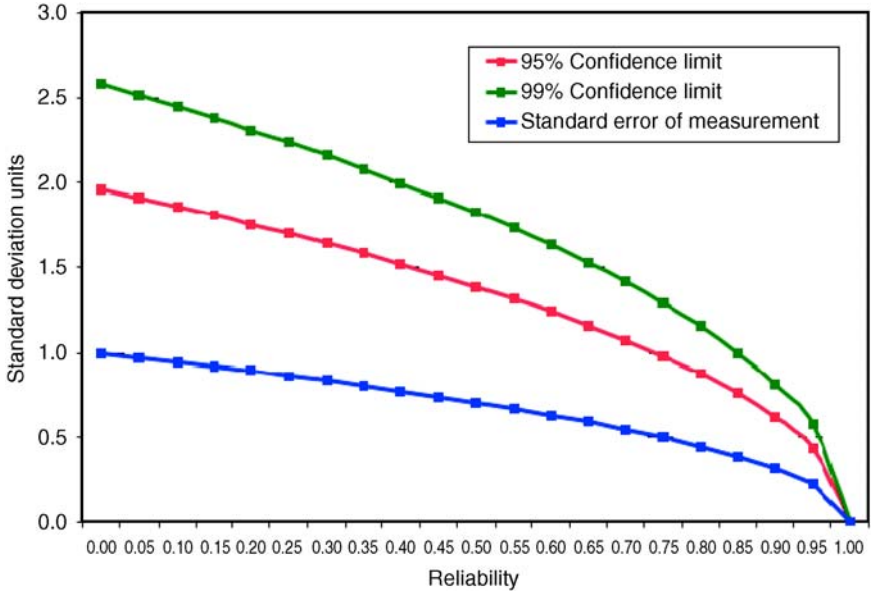

Figure 2 Standard errors of measurement and confidence limits as a function of scale reliability.

by five *T*-score points, the confidence intervals around that score are so large as to preclude such an interpretation. Clearly, results such as these should give us pause. Unfortunately, *SEM*s rarely are reported in test manuals for personality measures, despite the requirement for such in the *Standards for Educational and Psychological Testing* (American Psychological Association [APA], 1999).

Hence, for what level of reliability should we strive when building new measures? Opinions vary considerably on this question. As described earlier, high reliability is not always a good thing, especially when it is achieved by narrowing the scale with redundant items. Thus, some have argued that 0.80 is an acceptable level to shoot for in scale construction (e.g., Clark & Watson, 1995) and in basic research settings (Nunnally, 1978). However, in applied settings, where scores will be used to make important decisions about people, higher levels of internal consistency – as high as 0.90 or 0.95 – are recommended to ensure reasonably narrow standard errors around scores (Nunnally, 1978). As such, the prospective scale developer must consider carefully the amount of internal consistency necessary for the scale's intended purpose and then select scale items that accordingly balance measurement precision and proper scale breadth.

Time sampling error (i.e., test–retest reliability) also should be evaluated in this phase to the ensure that the scale is as stable as would be expected from one's theory of the construct. Trait personality scales, for example, should evidence greater temporal stability than state mood measures or

measures of more transient attitudes (e.g., Conley, 1984). Note, however, that test–retest correlations may be influenced by true change in the constructs being measured as well as memory or practice effects if the retest interval is too short. The potential impact of these factors should be carefully considered when planning a test–retest study of a new measure.

Scale homogeneity also should be considered in the scale construction process. Because internal consistency estimates such as coefficient alpha confound internal coherence with scale length, scale developers often use a variety of alternative approaches – including examination of interitem correlations (Clark & Watson, 1995) and conducting confirmatory factor analyses to test the fit of a single-factor model (Schmitt, 1996) – to assess the homogeneity of an item pool. To establish homogeneity using interitem correlations, one must examine both the average and distribution of the interitem correlations. According to Clark and Watson, the average interitem correlation should fall somewhere between 0.15 and 0.50, depending on the theoretical breadth (closer to 0.15) or narrowness (closer to 0.50) of the construct one is trying to measure. In addition, the distribution of interitem correlations should be inspected to ensure that they cluster narrowly around the average, since wide variation among the interitem correlations suggests a number of potential problems. For instance, excessively high correlations between certain pairs of items suggest unnecessary redundancy that can be eliminated by dropping one item from each such pair. Moreover, significant variability in the interitem correlations may be due to multidimensionality within the scale that must be explored (Cortina, 1993).

Interestingly, some scale developers faced with internal consistency problems will attempt to fix the problem by adding several highly redundant items to the scale. As described earlier, although this approach will increase alpha, it likely will do so at the expense of proper scale breadth and attenuate the scale's validity. A better approach when trying to increase scale reliability is to identify new items from the pool whose interitem correlations are close to the average interitem correlation. Doing so will increase internal consistency while maintaining the scale's homogeneity and balance between breadth and narrowness.

The role of modern psychometric theory in item selection

In recent years, scale developers have begun using a range of modern psychometric methods – mostly based on IRT – alone and as an adjunct to the classical methods previously discussed to inform item selection and scale evaluation. IRT refers to a range of models that describe the relations between item responses and the underlying latent trait they purport to measure. A complete account of IRT is beyond the scope of this paper; interested readers should see Embretson and Reise (2000), which is a reasonably accessible volume devoted to the use of IRT in

 Social and Personality Psychology Compass 2/1 (2008): 414–433, 10.1111/j.1751-9004.2007.00044.x

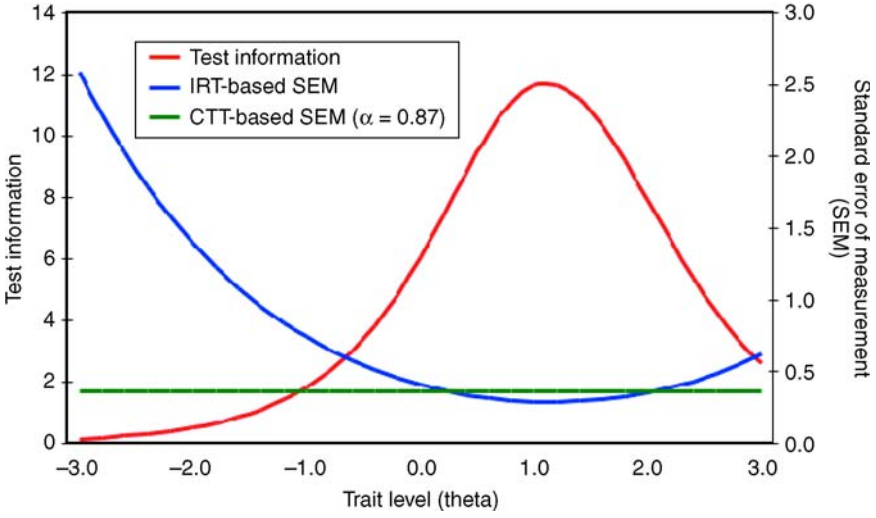

Figure 3 Test information and SEM curves for the Aggression scale of the Schedule for Nonadaptive and Adaptive Personality.

psychology. A variety of one-, two-, and three-parameter models have been proposed to explain both dichotomous and polytomous response data. Of these, a two-parameter model – with parameters for *item difficulty/severity* and *item discrimination* – has been applied most consistently to personality and psychopathology measures.

These item parameters can be combined to form an *item information curve*, which specifies where along the trait continuum a particular item provides the greatest measurement precision. In general, item difficulty determines the horizontal location of the curve's peak, and item discrimination influences the relative height of the peak compared to other items on the scale. Individual item information curves then can be summed to form a test information curve, which provides an overall index of measurement precision for a given scale.

Figure 3 includes an example of a test information curve that was computed for the aggression scale of the Schedule for Nonadaptive and Adaptive Personality (Clark, 1993) using the responses of 3,995 individuals who completed the Schedule for Nonadaptive and Adaptive Personality across a variety of samples. In contrast to the classical test theory methods described earlier – in which a constant level of precision is assumed across the entire range of a scale – the IRT concept of test information allows for conditional precision estimates at different levels of the underlying trait. In Figure 3, note how test information (marked by a red curve) peaks roughly at a trait level of 1.0 (interpreted similar to a *Z*-score metric), showing that this particular measure of aggression is most discriminating for individuals who are moderately high on this trait. Similarly, Figure 3

shows that the IRT-based *SEM* – that is equal to the inverse square root of information at every point along the trait continuum – is smallest at the peak of the test information curve and largest where the information curve is low. Compared to the classical test theory *SEM* (plotted in green, based on an alpha of 0.87 in this sample), it is clear that the IRT-based curves provide a more nuanced view of the scale's measurement properties.

Hence, how is one to use IRT techniques in the scale construction process? IRT clearly is relevant to the structural validity phase of scale development. For example, IRT-based information curves can be used to evaluate whether a new or existing scale includes adequate psychometric information at all levels of the trait that need to be reliably assessed. To that end, the aggression scale described above appears to be ideal for settings in which the goal is to discriminate between individuals who are moderate and high in trait aggression (such as in prison or psychiatric settings), but such a curve would not be ideal if one wishes to measure aggression with uniform precision across all levels of the trait dimension (which may be more important in epidemiological or community studies of personality). In the latter case, additional items would need to be written and evaluated that provide information for those below the midpoint on the trait as well as those very high on the trait.

Interestingly, despite the potential benefits associated with using IRT methods to build psychological scales, nearly all extant IRT applications relevant to social and personality psychology have focused on refining existing measures rather than building new measures from scratch. This is regrettable, as post hoc attempts to fix problematic scales with IRT likely will not be as successful as building new scales using modern methods. Part of the problem may be that IRT methods are not routinely discussed in texts and courses that teach about psychological scale construction. Hence, how would one go about building an IRT-based scale from scratch? This is a large question that goes beyond the scope of this paper, but I will discuss several possible ways to improve scale development through the use of IRT. First, IRT methods can be used to confirm the dimensionality of a given item pool and to select items that (i) discriminate above a reasonable level, and (ii) represent all levels of the dimension that the scale developer wishes to measure reliably. Just as with the aggression scale example above, examination of item and test information curves can be extremely useful in this regard. If, for example, the initial IRT analyses for a given new scale reveal that the scale lacks enough discriminating items in an important region of the trait dimension, then additional items can be written and tested prior to finalizing the scale.

The IRT methods also can be used to identify biased items, or *differential item functioning* (DIF). Such 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). Ideally, DIF analyses

should be done during the scale construction process – as opposed to in a post hoc manner – to identify and fix all problematic items. Finally, IRT methods are very useful for creating computerized adaptive tests (CATs), in which computers tailor measures to respondents by iteratively selecting and administering items that provide the most 'psychometric information' given the respondent's trait level. Compared to using traditional measures, properly built CATs generally yield significant item savings and equivalent reliability and validity. Although primarily used in the educational testing literature, CAT methods recently have shown utility in the personality literature (e.g., Reise & Henson, 2000; Simms & Clark, 2005).

External Validity Phase

The final phase of scale construction is the external validity phase in which relations between the new measure and important test and nontest criteria are studied to determine whether they are congruent with one's theoretical understanding of the target construct and its position with respect to other similar and dissimilar constructs – what Cronbach and Meehl (1955) termed the nomological net. Validity evidence can take many forms, such as correlations with other measures as well as nontest criteria relevant to the construct's nomological net. Correlations that are consistent with one's theory of the construct support the construct validity of the new measure. Discrepancies between observed results and one's theory suggest that (i) the measure does not adequately measure the target construct, (ii) the theory requires modification, or (iii) some combination of both.

External validity studies involve assessment of several related aspects of construct validity: (i) convergent and discriminant validity, and (ii) concurrent and predictive validity (known collectively as criterion-related validity). *Convergent validity* is the extent to which a measure correlates with other indicators of the same construct, whereas *discriminant validity* is the extent to which a measure does not correlate with indicators of other constructs that are theoretically or empirically distinct. *Concurrent validity* involves relating a measure to criteria assessed at the same time as the measure itself, whereas *predictive validity* involves associations with criteria that are assessed at some point in the future. Unfortunately these various types of validity evidence often are confused in the literature and applied inconsistently. Thus, how is one to reconcile their similarities and differences? Rather than thinking of them as independent types of validity, it is useful to consider them as different aspects of the same validity evidence. For example, pretend that we wish to validate a new measure of extraversion. As depicted in Figure 4, the same piece of validity evidence for this measure (e.g., a wife's current rating of her husband's extraversion) can contribute to both the convergent validity and concurrent validity of the measure. Similarly, another piece of evidence (e.g., ratings of academic

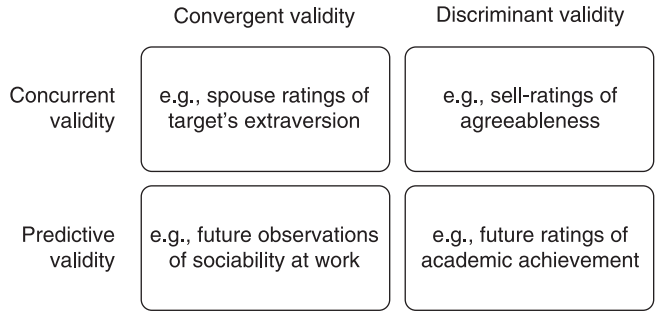

Figure 4 Examples of different types of construct validity evidence.

Table 2 Example of a hypothetical multitrait–multimethod matrix

		Method A			Method B		
Method	**Trait**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**
Method A	T1	(0.85)					
	T2	0.25	(0.86)				
	T3	0.22	0.20	(0.83)			
Method B	T1	<u>0.45</u>	0.15	0.10	(0.91)		
	T2	0.09	<u>0.51</u>	0.05	0.30	(0.93)	
	T3	0.17	0.18	<u>0.42</u>	0.25	0.27	(0.90)

Note: Alpha coefficients are presented in parentheses along the diagonal. Convergent correlations are underlined.

achievement obtained at some point in the future) could be used as evidence of the new measure's predictive and discriminant validity, since we would have little theoretical reason to expect extraversion to predict future academic performance.

Campbell and Fiske (1959) argued that these various aspects of construct validity should be assessed using a multitrait–multimethod (MTMM) matrix. In such a matrix, the intercorrelations among multiple measures of at least two constructs are computed and arranged to highlight several important aspects of convergent and discriminant validity. Ideally, to account for the influence of shared method variance, data reflecting at least two different measurement methods (e.g., self-report, peer-report, behavioral observation) should be included in such a matrix. To illustrate the relevant aspects of a MTMM matrix, a hypothetical example appears in Table 2. In this example, intercorrelations are presented among three traits (T1, T2, and T3) assessed using two different methods (A and B). The numbers in parentheses along the top diagonal are reliability coefficients;

they generally are the largest coefficients in the matrix. The underlined values along the diagonal in the lower-left box represent convergent validity coefficients comparing the same traits (monotrait) across different methods (heteromethod). These should be positive and at least moderate in size.

An ideal MTMM matrix includes convergent correlations that are greater than all other correlations in the table, thereby establishing discriminant validity. Three specific comparisons are made to establish discriminant validity. First, Campbell and Fiske (1959) recommended that each convergent correlation should be higher than other correlations in the same row and column in the same box (*heterotrait–heteromethod correlations*). Second, Campbell and Fiske argued that the convergent correlations should be higher than the correlations in the *heterotrait-monomethod triangles* that appear above and to the right of the heteromethod block just described. Finally, they suggested that 'the same pattern of trait interrelationship [should] be shown in all of the heterotrait triangles' (Campbell & Fiske, 1959, 83).

The values presented in this hypothetical example appear to satisfy all of Campbell and Fiske's recommendation for good construct validity. However, in real-world applications, these predicted relations will not always be so clean and should be evaluated using statistical tests designed to detect differences between correlations (see Steiger, 1980). Construct validity is supported to the extent that a new measure manifests this predicted pattern of relations with appropriately chosen measures of similar and dissimilar constructs assessed with multiple methods. Convergent and discriminant validity can be quantified in other ways as well. For example, using confirmatory factor analysis, observed variables can be modeled to load both on trait and method factors, thereby allowing for the relative influence of each to be quantified. Also, some have tried to reduce construct validity down to a smaller number of coefficients that can then be compared across measures (e.g., Smith, 2005; Westen & Rosenthal, 2003).

Finishing Up

Once sufficient reliability and validity data have been collected to support the internal structure and construct validity of a new measure, the scales should be finalized. In addition, as specified by the *Standards* (APA, 1999), scale developers should produce a research article or test manual that thoroughly describes the methods used to construct the measure, appropriate administration and scoring procedures, and interpretive guidelines (APA, 1999).

In this paper, I present a model of psychological scale construction that is rooted in the classic works of Loevinger (1957), Cronbach and Meehl (1955), and Campbell and Fiske (1959). A general theme to be highlighted is that construct validity of a measure is not a static entity that can be 'established' in any definitive sense. Rather, construct validation is a

dynamic process in which theory informs the scale development process at all phases, and results of the scale development have the potential to modify our theoretical understanding of the target construct. In addition to methods rooted in classical test theory, I present several principles of IRT and discuss how such methods can be used to help evaluate and select items in the structural phase of scale development. These new methods offer much to the scale developer, and it is hoped that IRT will play a more prominent role in scale development as the techniques are more widely disseminated.

Acknowledgment

The author thanks David Watson for his comments regarding a specific part of this paper.

Short Biography

Leonard Simms conducts research and publishes papers that are broadly relevant to measurement of and theory related to personality and psychopathology. More specifically, he studies applied and basic psychological assessment, dimensional models of personality and psychopathology, item response theory applications to personality measurement, and computerized adaptive testing. He is active in a number of professional societies and serves on the editorial boards for *Journal of Abnormal Psychology* and *Assessment*. Dr. Simms currently is on the faculty at the University at Buffalo, the State University of New York. He holds a BS in Psychology from California Polytechnic State University and MA and PhD degrees in Clinical Psychology from the University of Iowa.

Endnote

* Correspondence address: Department of Psychology, Park Hall 218, University at Buffalo, The State University of New York, Buffalo, NY 14260, USA. Email: ljsimms@buffalo.edu.

References

American Psychological Association (1999). *Standards for Educational and Psychological Testing*. Washington, DC: Author.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). New York: Macmillan.

Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist*, **39**, 214–227.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory (MMPI-2). Manual for Administration and Scoring*. Minneapolis, MN: University of Minnesota Press.

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

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

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, **7**, 309–319.
Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, **56**, 754–761.
Conley, J. J. (1984). The hierarchy of consistency: A review and model of longitudinal findings on adult individual differences in intelligence, personality, and self-opinion. *Personality and Individual Differences*, **5**, 11–25.
Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, **78**, 98–104.
Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, **52**, 281–302.
Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, **4**, 272–299.
Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, **7**, 286–299.
Gough, H. G. (1987). *California Psychological Inventory Administrator's Guide*. Palo Alto, CA: Consulting Psychologists Press.
Gulliksen, H. (1950). *Theory of Mental Tests*. New York: John Wiley & Sons.
Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.
Hathaway, S. R., & McKinley, J. C. (1943). *Manual for the Minnesota Multiphasic Personality Inventory*. New York: Psychological Corporation.
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.
Huang, C., Church, A., & Katigbak, M. (1997). Identifying cultural differences in items and traits: Differential item functioning in the NEO Personality Inventory. *Journal of Cross Cultural Psychology*, **28**, 192–218.
John, O. P., & Soto, C. J. (2007). The importance of being valid: Reliability and the process of construct validation. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology* (pp. 461–494). New York: Guilford.
Kaplan, R. M., & Saccuzzo, D. P. (2005). *Psychological Testing: Principles, Applications, and Issues* (6th ed.). Belmont, CA: Thomson Wadsworth.
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.
Mackinnon, A., Jorm, A. F., Christensen, H., Scott, L. R., Henderson, A. S., & Korten, A. E. (1995). A latent trait analysis of the Eysenck Personality Questionnaire in an elderly community sample. *Personality and Individual Differences*, **18**, 739–747.
Meehl, P. E. (1945). The dynamics of structured personality tests. *Journal of Clinical Psychology*, **1**, 296–303.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, **50**, 741–749.
Nunnally, J. C. (1978). *Psychometric Theory*. New York: McGraw-Hill.
Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, **2**, 13–43.
Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO-PI-R. *Assessment*, **7**, 347–364.
Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, **8**, 350–353.
Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality. *Psychological Assessment*, **17**, 28–43.
Simms, L. J., & Watson, D. (2007). The construct validation approach to personality scale construction. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology* (pp. 240–258). New York: Guilford.

Smith, G. T. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment*, **17**, 396–408.

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction Scale. *Journal of Personality and Social Psychology*, **75**, 1350–1362.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, **87**, 245–251.

Watson, D. (2006). In search of construct validity: Using basic concepts and principles of psychological measurement to define child maltreatment. In M. Feerick, J. Knutson, P. Trickett, & S. Flanzer (Eds.), *Defining and Classifying Child Abuse and Neglect for Research Purposes*. Baltimore, MD: Brookes Publishing.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, **84**, 608–618.

