Published in final edited form as:
Psychol Assess. 2019 December ; 31(12): 1412–1427. doi:10.1037/pas0000626.

Constructing Validity: New Developments in Creating Objective Measuring Instruments

Lee Anna Clark, **David Watson [Psychological Assessment]**
Department of Psychology, University of Notre Dame

Abstract

In this update of Clark and Watson (1995), we provide a synopsis of major points of our earlier article and discuss issues in scale construction that have become more salient as clinical and personality assessment has progressed over the past quarter-century. The primary goal of scale development is still to create valid measures of underlying constructs and that Loevinger's theoretical scheme provides a powerful model for scale development. We still discuss practical issues to help developers maximize their measures' construct validity, reiterating the importance of (1) clear conceptualization of target constructs; (2) an overinclusive initial item pool; (3) paying careful attention to item wording; (3) testing the item pool against closely related constructs; (4) choosing validation samples thoughtfully; and (5) emphasizing unidimensionality over internal consistency. We have added (6) consideration of the hierarchical structures of personality and psychopathology in scale development; discussion of (7) co-developing scales in the context of these structures; (8) "orphan," and "interstitial" constructs, which do not fit neatly within these structures; (9) problems with "conglomerate" constructs; and (10) developing alternative versions of measures, including short forms, translations, informant versions, and age-based adaptations. Finally, we have expanded our discussions of (10) item-response theory and of external validity, emphasizing (11) convergent and discriminant validity, (12) incremental validity, and (13) cross-method analyses, such as questionnaires and interviews. We conclude by re-affirming that all mature sciences are built on the bedrock of sound measurement and that psychology must redouble its efforts to develop reliable and valid measures.

Keywords

test construction; scale development; questionnaire development; measurement, construct validity; convergent validity; discriminant validity; incremental validity; item analysis; psychopathology; personality structure

Clark and Watson (1995) discussed theoretical principles, practical issues, and pragmatic decisions in the process of objective scale development to maximize the construct validity of measures. In this article, we reiterate a few points we discussed previously that continue to

There is no licensing fee for use of the SNAP family of measures for non-commercial, unfunded research. For all other uses, LAC negotiates a mutually acceptable licensing fee.

Correspondence concerning this article should be addressed to Lee Anna Clark, Department of Psychology, University of Notre Dame, E390 Corbett Family Hall, Notre Dame, IN 46556. may be sent to la.clark@nd.edu.
Lee Anna Clark and David Watson, Department of Psychology, The University of Notre Dame.

be underappreciated or largely ignored, but we primarily discuss additional issues that have become important due to advances in the field over the past 2+ decades. As before, we focus on language-mediated measures (vs., e.g., coding of direct observations), on scales with clinical relevance (i.e., those of most interest to readers of this journal) and on measures that are indeed intended to measure a construct (vs., e.g., a checklist to assess mortality risk that could be used to make valid inferences for the purpose of life insurance premiums[1]). Readers are encouraged to review the previous paper for points underdeveloped in this one, as we often provide here only a synopsis of the earlier material.

The Centrality of Psychological Measurement

Measurement is fundamental in science, and, arguably, the two most important qualities related to measurement are reliability and validity. Note that we say "measurement" not "measure." Despite the thousands of times that some variant of the phrase "(measure) X has been shown to have good reliability and validity" has appeared in articles' Method sections,[2] the phrase is vacuous. Validity in particular is not a property of a *measure*, but pertains to *interpretations of measurements*. As first stated by Cronbach and Meehl (1955), "One does not validate a test, but only a principle for making inferences" (p. 297). Similarly, the 5th edition of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) states unequivocally, "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of a test. Validity is, therefore, the most fundamental consideration in developing and evaluating tests" (p. 11). Accordingly, investigating a measure's construct validity necessarily involves empirical tests of hypothesized relations among theory-based constructs and their observable manifestations (Cronbach & Meehl, 1955), and absent an articulated theory (i.e., "the nomological net"), there is no construct validity. Despite the implication that a *series* of interrelated investigations is required to understand the construct(s) that a measure assesses, scale developers often speak rather lightly of establishing a scale's construct validity in its initial publication. Cronbach and Meehl (1955) was published 60-plus years ago, and 30+ years have passed since the 3rd edition of the *Standards* (APA, 1985), which firmly established construct validity as the core of measurement. Yet, there remains widespread misunderstanding regarding the overarching concept of construct validity and what establishing construct validity entails. Clearly, test developers, not to mention test users, either do not fully appreciate or willfully choose to ignore the complexity and importance of the concept.

Why should I care about construct validity?

First, construct validity is the foundation of clinical utility. That is, to the extent that real-world decisions (e.g., eligibility for social services, psycho- or pharmaco-therapy selection) are based on psychological measurements, the quality of those decisions depends on the construct validity of the measurements on which they are based. Second, practitioners increasingly are asked to justify use of specific assessment procedures to third-party payers.

[1] We thank an anonymous reviewer with pointing out this issue and providing this example.

[2] So as not to be completely hypocritical, we admit that we, too, use this short-hand language.

Use of psychological measures whose precision and efficiency are well established within an articulated theory that is well supported by multiple types of empirical data (i.e., measurements with demonstrated construct validity) may be required in the future. Third, progress in psychological science, especially as we explore more deeply the interface between psychosocial and neurobiological systems, is critically dependent on measurement validity. Detailed understanding of brain activity will be useful only insofar as we can connect it to phenotypic phenomena, so the more validly and reliably we can measure experienced affects, behaviors, and cognitions, the more we will be able to advance psychology *and* neuroscience.

A Theoretical Model for Scale Development

Loevinger's (1957) monograph remains the most complete exposition of theoretically based psychological test construction. In both the previous and this article, we offer practical guidance for applying Loevinger's theoretical approach to the process of scale development, with specific emphasis on the "three components of construct validity": substantive, structural, and external.

Substantive Validity: Conceptualization and Development of an Initial Item Pool

Conceptualization.—There is essentially no limit to the number of psychological constructs that can be operationalized as scales, and sometimes it seems that there is a scale for every human attribute (e.g., adaptability, belligerence, complexity, docility, efficiency, flexibility, grit, hardiness, imagination, . . .zest). However, not all of these represent a sufficiently important and distinct construct to justify scale development. As a thought experiment, imagine listing not only the thousands of such constructs in the English language (Allport & Odbert, 1936), but also doing the same for the approximately 7,000 existing human languages. No doubt many of the constructs these words represent are highly overlapping and it would be absurd to argue that each one would make a significant difference in predicting real-world outcomes. *This point holds true even within just the English language.* Thus, an essential early step is to crystallize one's conceptual model by writing a precise, reasonably detailed description of the target construct.

Literature Review.—To articulate the basic construct as clearly and thoroughly as possible, this step should be embedded in a literature review to ensure that the construct doesn't already have one or more well-constructed measures and to describe the construct in its full theoretical and hierarchical-structural context, including its level of abstraction and how it is distinguished from near-neighbor constructs. For instance, in developing a new measure of hopelessness, the literature review would encompass not only existing measures of hopelessness, but also measures of related, broader constructs (e.g., depression and optimism-pessimism), and somewhat less immediately related constructs that might correlate with the target construct, such as various measures of negative affect (anxiety, guilt and shame, dissatisfaction, etc.) in order to articulate the hypothesized overlap and distinctiveness of hopelessness in relation to other negative affects. That is, conceptual models must articulate both what a construct is and what it is not. The importance of a comprehensive literature review cannot be overstated, because it enables a clear articulation of how the proposed measure(s) will be either a theoretical or an empirical improvement over existing measures or fill an important measurement gap.

Our emphasis on theory and structure is not meant to intimidate or to imply that one must have from the outset a fully articulated set of interrelated theoretical concepts and know in advance exactly how the construct will fill a gap in an established hierarchy of psychological constructs or improve measurement over existing scales. Rather, our point is that serious consideration of theoretical and structural issues prior to scale construction increases the likelihood that the resulting scale will make a substantial contribution by providing significant incremental validity over existing measures, a topic we return to in a subsequent section.

Hierarchical structure of constructs.—It is now well established that psychological constructs—at least in the clinical and personality domains that are our focus—are ordered hierarchically at different levels of abstraction or breadth (see Comrey, 1988; Watson, Clark, & Harkness, 1994). In personality, for instance, one can conceive of the narrow-ish traits of “talkativeness” and “attention-seeking,” the somewhat broader concepts of “gregariousness” and “ascendance” that encompass these more specific terms, respectively, and the still more general disposition of “extraversion” that subsumes all these lower order constructs. Scales can be developed to assess constructs at each level of abstraction, so a key initial issue that is too often overlooked is the *level* at which a construct is expected to fit in a particular structure.

As our knowledge of the hierarchical structure of personality and psychopathology has grown, so too has the importance of considering measures as elements in these structures (vs. focusing on developing a single, isolated scale). In particular, the broader the construct (i.e., the higher it lies in the hierarchy), the more important it is to articulate its lower level components—that is, to explicate the nature of its multidimensionality. This has become important enough that in many cases, new scales should not be developed in isolation, but rather should be “co-developed” with the express intent of considering their convergent and discriminant validity as part of initial scale development (i.e., not leaving consideration of convergent and discriminant validity to the later external-validation phase). We discuss this further in a subsequent section.

Orphan, interstitial, and conglomerate constructs.: Growth in our understanding of hierarchical structures has increased awareness that not all constructs fit neatly into these structures. We consider three types of such constructs. *Orphan* constructs are unidimensional constructs that load only weakly on any superordinate dimension, and their value is relative to their purpose: They may have significant incremental predictive power over established constructs for specific outcomes or be important in a particular area of study. For example, intrinsic religiosity is largely unrelated to the personality-trait hierarchy (Lee, Ogunfowora, & Ashton, 2005), yet it predicts various mental-health outcomes across the lifespan (e.g., Ahmed, Fowler, & Toro, 2011). Similarly, dependency is an important clinical construct t does not load strongly on any primary personality domain (e.g., Lowe, Edmundson, & Widiger, 2009).

Interstitial constructs are both unidimensional and blends of two distinct constructs, such that factor analysis of their items yields (1) a single factor on which all scale items load and/or (2) two (or more) factors that are either (a) orthogonal with all or almost all items loading on both (all) factors, or (b) highly correlated (i.e., per an oblique rotation). For example, Watson, Suls, & Haig (2002) showed that self-esteem is unidimensional but yet has strong loadings on both (low) negative affectivity and positive affectivity, because the measure's items themselves inherently encompass variance from both higher order dimensions. Some interstitial constructs blend two dimensions at the same hierarchical level: In the interpersonal circumplex (IPC; Zimmermann & Wright, 2017), dominance and affiliation typically constitute the primary axes, and dimensions that fall between these define interstitial constructs (e.g., IPC arrogance and introversion-extraversion). In a perfect circumplex, the axes location is arbitrary; in reality, psychological theory and the utility of the measures' empirical findings guide axis placement.

Finally, *conglomerate* constructs are intended to be "winning combinations" of two or more modestly to moderately related constructs. For example, the popular construct *grit*, defined as "perseverance and passion for long-term goals" (Duckworth, Peterson, Matthews, & Kelly, 2007, p. 1087) was intended to predict success in domains as variant as the National Spelling Bee and West Point. A recent meta-analysis, however, found that its perseverance facet better predicted success than the construct as a whole (Credé, Tynan, & Harms, 2017), thus challenging the theoretical basis of the construct. More generally, conglomerate constructs rarely fulfill their enticing premise that the total is greater than the sum of its parts. If development of such a construct is pursued, the burden of proof is on the developer to show that the conglomerate is superior to the linear combination of its components.

We stress here that we are ***not*** suggesting that one should seek to eliminate these types of constructs from one's measurement model and strive for all scales to mark a single factor clearly. On the contrary, orphan and interstitial constructs—with low and cross-loadings, respectively—are particularly important for providing a full characterization and understanding of hierarchical structures of personality and psychopathology. We aim rather to alert structural researchers to the fact that this variance also should be recognized and appropriately modeled.

Broader implications of hierarchical models.: The emergence of hierarchical models has led to the important recognition that scales—even highly homogeneous ones—contain multiple sources of variance that reflect different hierarchical levels. For example, a well-designed assertiveness scale contains not only construct-specific variance reflecting stable individual differences in this lower order trait, but also includes shared variance with other lower order components of extraversion (e.g., gregariousness and positive emotionality; Watson, Stasik, Ellickson-Larew, & Stanton, 2015), which reflects the higher order construct of extraversion. Thus, a lower order scale simultaneously contains both unique (lower order facet) and shared (higher order trait) components. Multivariate techniques such as multiple regression (e.g., Watson, Clark, Chmielewski, & Kotov, 2013) and bifactor analysis (e.g., Mansolf & Reise, 2016) can be used to isolate the specific influences of these different elements.

It is less widely recognized that *items* also simultaneously contain multiple sources of variance reflecting different levels of the hierarchy in which they are embedded. We illustrate this point using data from a large sample (*N* = 8,305) that includes patients, adults, postpartum women, and college students (Watson et al., 2013, Study 1). All participants completed the Inventory of Depression and Anxiety Symptoms (IDAS; Watson et al., 2007).

We focus on the IDAS item *I woke up much earlier than usual.* At its most specific level, this item can be viewed as an indicator of the construct of terminal insomnia/ early morning awakening: It correlates strongly with another terminal-insomnia item (*r* = .65 with *I woke up early and could not get back to sleep*) and could be used to create a very narrow measure of this construct. However, this item also correlates moderately (*r*s = .34 to .45; see the upper portion of Table 1) with items assessing other types of insomnia and could be combined with those items to create a unidimensional measure of insomnia: We subjected these items to a confirmatory factor analysis (CFA) using PROC CALIS in SAS 9.4 (SAS Institute, Inc., 2013) to test how well a single factor modeled their intercorrelations. We used four fit indices to evaluate the model: the standardized root-mean-square residual (SRMR), root-mean-square error of approximation (RMSEA), comparative fit index (CFI), and Tucker-Lewis Index (TLI).[3] A 1-factor model fit these data extremely well (CFI = .996, TLI = .988, SRMR = .013, RMSEA = .048), demonstrating that this item is a valid indicator of general insomnia.

Further, this item correlates moderately (*rs* = .22 to .26; middle portion of Table 1) with other symptoms of major depression and could be used to create a unidimensional measure of this broader construct. A CFA of these items also indicated that a 1-factor model fit the data well (CFI = .983, TLI = .948, SRMR = .023, RMSEA = .068). Thus, this item *also* is a valid indicator of depression. Finally, at an even broader level of generality, this item is moderately related (all *r*s ≥ .24; bottom portion of Table 1) to indicators of internalizing psychopathology and could be combined with them to create a unifactorial measure of this overarching construct. Again, a CFA of these items indicated that a 1-factor model fit the data very well (CFI =.998, TFI = .995, SRMR = .007, RMSEA = .018), showing that this item is *also* a valid indicator of internalizing.

In theory, one could extend this analysis to establish that this item also reflects the influence of the general factor of psychopathology (Tackett et al., 2013). Consequently, structural analyses—based on different sets of indicators reflecting varying hierarchical levels—could be used to establish that the item *I woke up much earlier than usual* is simultaneously an indicator of (a) terminal insomnia, (b) general insomnia, (c) depression, (d) internalizing, and (e) general psychopathology. Note, moreover, that this complexity is *inherent in the item itself.* We could have used any number of items to illustrate this point, so these analyses strongly support the assertion that items and scales typically reflect multiple meanings and constructs, not simply a single set of inferences (*Standards,* p. 11). They also highlight the importance of writing and selecting items very carefully during the process of scale construction.

[3] Fit is generally considered acceptable if CFI and TLI are .90 or greater and SRMR and RMSEA are .10 or less (Finch & West, 1997; Hu & Bentler, 1998); and as excellent if CFI and TLI are .95 or greater and SRMR and RMSEA are .06 or less (Hu & Bentler, 1999).

Creation of an item pool.—The next step is item writing. No existing data-analytic technique can remedy item-pool deficiencies, so this is a crucial stage whose fundamental goal is to sample systematically all *potentially* relevant content to ensure the measure's ultimate content validity. Loevinger (1957) offered the classic articulation of this principle: "*. . .the items of the pool should be chosen so as to sample all possible contents which might comprise the putative trait according to all known alternative theories of the trait"* (p. 659; emphasis in original). Two key implications of this principle are: the initial pool should be broader and more comprehensive than one's theoretical view of the target construct and include content that ultimately will be eliminated. Simply put, psychometric analyses can identify items to drop but not missing content that should have been included; accordingly, one initially should be overinclusive.

In addition, the item pool must include an adequate sample of each major content area that potentially composes the construct, because undersampled areas likely will be underrepresented in the final scale. To ensure that all aspects of a construct are assessed adequately, some test developers recommend creating formal subscales, called homogeneous item composites (Hogan, 1983) or factored homogeneous item dimensions (Comrey, 1988), to assess each content area (see Watson et al., 2007, for an example). Ideally, the number of items in each content area should be proportional to that area's importance in the target construct, but often the theoretically ideal proportions are unknown. In general, however, broader content areas should be represented by more items than narrower ones.

Many of these procedures are traditionally described as reflecting the *theoretical-rational* or *deductive* method of scale development (e.g., Burisch, 1984), but we consider them an initial step in an extensive process, not a "stand-alone" scale-development method. Loevinger (1957) emphasized that attending to content was necessary, but not sufficient; rather, empirical validation of content was critical: "If theory is fully to profit from test construction..., every item [on a scale] must be accounted for" (Loevinger, 1957, p. 657). This obviously is an ideal to be striven for, not an absolute requirement (see also Comrey, 1988; and Haynes, 1995).

Good scale construction is an iterative process involving several stages of item writing, each followed by conceptual and psychometric analysis that sharpen one's understanding of the nature and structure of the target domain and may identify shortcomings in the initial item pool. For instance, factor analysis might identify subscales and also show that the initial pool contains too few items to assess one or more content domains reliably. Accordingly, new items must be written, and additional data collected and analyzed. Alternatively, analyses may suggest that the target construct's original conceptualization is countermanded by the empirical results, requiring revision of the theoretical model, a point we develop further later.

Basic principles of item writing.—In addition to sampling well, it is essential to write "good" items, and it is worth the time to consult the item-writing literature on how to do this (e.g., Angleitner & Wiggins, 1985; Comrey, 1988). We mention only a few basic principles here. Items should be simple, straightforward, and appropriate for the target population's reading level. Avoid (1) expressions that may become dated quickly; (2) colloquialisms that may be not be familiar across age, ethnicity, region, gender, and so forth; (3) items that virtually everyone (e.g., "Sometimes I am happier than at other times") or no one (e.g., "I am always furious") will endorse; and (4) complex or "double-barreled" items that assess more than one characteristic; for example, "I would never drink and drive for fear that I might be stopped by the police," assesses both a behavior's (non)occurrence and a putative motive. Finally, the exact phrasing of items can greatly influence the construct that is being measured. For example, the inclusion of almost any negative mood term (e.g., "I worry about...," I am upset [or bothered or troubled] by...") virtually guarantees a substantial neuroticism/ negative affectivity component to an item.

Choice of format

Two dominant formats.—Currently, the two dominant response formats in personality and clinical assessment are dichotomous responding (e.g., true/false; yes/no) and Likert-type rating scales with three or more options. There are several considerations in choosing between these, but surprisingly little empirical research. Recently, however, Simms, Zelazny, Williams, and Bernstein (2019) systematically examined an agree–disagree format with response options ranging from 2 to 11, evaluating the psychometric properties and convergent validity of a well-known personality trait measure in a large undergraduate sample. Their results indicated that psychometric quality (e.g., internal consistency reliability, dependability) increased up to six response options, but the number of response options had less effect on validity than expected.

Likert-type scales are used with various response formats, including frequency (e.g., "never" to "always"), degree or extent (e.g., "not at all" to "very much"), similarity (e.g., "very much like me" to "not at all like me"), and agreement (e.g., "strongly agree" to "strongly disagree"). Obviously, the nature of the response format constrains item content in an important way and vice versa (Comrey, 1988). For example, frequency formats are inappropriate if the items themselves use frequency terms (e.g., "I often lose my temper"). Whether to label all or only some response options also must be decided. Most measures label up to about six response options; beyond which they vary (e.g., only the extremes, every other response, etc.). With an odd number of response options, the middle option's label must be considered carefully (e.g., "cannot say" confounds uncertainty with a mid-range rating such as "neither agree nor disagree"), whereas even numbers of response options force respondents to "fall on one side of the fence or the other," which some respondents dislike. However, Simms et al. found no systematic differences between odd versus even number of response options. More research of this type is needed using a broad range of constructs (e.g., psychopathology, attitudes), samples (e.g., patients, community adults), type of response formats (i.e., extent, frequency), and so on.

Less common formats.—Checklists have fallen out of favor because they are more prone to response biases (e.g., D. P. Green, Goldman, & Salovey, 1993), whereas visual analog scales—now easily scored via computer administration—are making a come-back, particularly in studies of medical problems using simple ratings of single mood terms or problem severity (e.g., pain, loudness of tinnitus). Simms et al. (2019) found them to be only slightly less reliable/ valid than numerical rating scales. Forced-choice formats, which are largely limited to legacy measures in personality and clinical assessment, also are making a come-back in the industrial-organizational psychology literature due to advances in statistical modeling techniques that solve problems of ipsative data that previously plagued this format (e.g., Brown & Maydeu-Olivares, 2013). The advantage of this format is reduction in the effect of social desirability responding, but unfortunately, we do not have the space to do justice to these developments here.

Derivative versions

Different versions of a measure may be needed for a range of purposes and include short forms, translations, age-group adaptations, and other-report forms (e.g., for parents, spouses). Adaptations require re-validation, but far too often this step is skipped or given short shrift. We have space only to raise briefly some key issues in derivations. When available, we refer readers to sources that provide more detailed guidance for their development.

Short forms.—Smith, McCarthy, and Anderson (2000) provide an excellent summary of the many challenges in developing and validating short forms. They address the tendency to try to maintain a similar level of internal consistency (e.g., coefficient alpha) by narrowing the content, which leads into the classic *attenuation paradox* of psychometrics (Boyle, 1991; Loevinger, 1954): Increasing a test's internal consistency beyond a certain point can reduce its validity relative to its initially intended interpretation(s). Specifically, by narrowing the scale content, the scope and nature of the assessed construct is itself changed; in particular, it increases item redundancy, thereby reducing the total amount of construct-related information the test provides.

Broadly speaking, the central challenge in creating short forms is to maintain the level of test information while simultaneously significantly reducing scale length. Analyses based on item response theory (IRT; Reise, Ainsworth, & Haviland, 2005; Simms & Watson, 2007) can be invaluable for this purpose by providing a detailed summary of the nature of the construct-related information that each item provides, which can be used to identify a reduced set of items that yields maximal information. Thus, we strongly support the increasing use of IRT to create short forms (e.g., Carmona-Perera, Caracuel, Pérez-García, & Verdejo-García, 2015).

Developing short forms also provides an opportunity to improve a measure's psychometric properties, particularly in the case of hierarchical instruments. For example, some of the domain scores of the widely used Revised NEO Personality Inventory (NEO PI-R; Costa, & McCrae, 1992) correlate substantially with one another, which lessens their discriminant validity (e.g., Costa & McCrae, 1992, reported a −.53 correlation between Neuroticism and Conscientiousness). The measure contains six lower order facet scales to assess each domain, but in creating a short form—the NEO Five-Factor Inventory (NEO-FFI; Costa & McCrae, 1992)—the authors selected the best markers of each higher order domain, rather than sampling equally across their facets, which improved the measure's discriminant validity. In a student sample ($N = 329$; Watson, Clark, & Chmielewski, 2008), the NEO-FFI

Neuroticism and Conscientiousness scales correlated significantly lower ($r = -.23$) than did those of the full NEO PI-R version ($r = -.37$).

As new technologies have enabled new research methods, such as ecological momentary assessment, in which behavior is sampled multiple times over the course of several days or weeks, and as researchers increasingly investigate complex interplays of diverse factors influencing such outcomes as psychopathology or crime via multi-level modeling, the demand for very short forms—one to three items—has increased. These approaches offer important new perspectives on human behavior, but meaningful and generalizable results still depend on measures' reliability and validity. Psychometric research is accruing on measurement relevant to these methods. For example, Soto and John (2017) reported that their 15-item extra-short form (of the 60-item Big Five Inventory-2), which has one item per facet, should be used only to derive domain and not facet scores, whereas the 30-item short form could be used to assess facets. Similarly, McCrae (2014) cautioned that, on average, only one-third of single items' variance reflected the target construct for broad traits (e.g., neuroticism); moreover, items may capture valid sub-facet or "nuance" variance, but little is known about the construct validity of nuances compared to domains and facets. The crucial point is that it remains incumbent on the researcher to demonstrate the validity of the inferences drawn from ultra-short form measures.

Translations.—Translations into western European languages may be the "worst offenders" in terms of ignoring revalidation, in that articles often do not even indicate that study measures are translations other than by implication (e.g., stating that the data were collected in Germany). There is quite a large literature on translation validation to which we cannot do justice here, so we simply refer readers to a chapter that explicitly discusses the key issues (Geisinger, 2003) and two examples of strong translation development and validation processes (Schwartz et al., 2014; Watson, Clark, & Tellegen, 1984). This issue is sufficiently important that the *Test Standards* explicitly state that both those who develop and use translations/ adaptations are responsible for providing evidence of their construct validity (AERA, APA, & NCME, 2014; Standard 3.12).

Informant versions.—There are various reasons for collecting information from individuals other than, or in addition to, target individuals: difficulty or inability in responding for themselves (e.g., individuals with dementia, McDade-Montez, Watson, O'Hara, & Denburg, 2008; or children, Putnam, Rothbart, & Gartstein, 2008); concerns about valid responding (e.g., psychopathology, Achenbach, Krukowski, Dumenci, & Ivanova, 2005), and simply to provide another perspective on the person's behavior (e.g., Funder, 2012).

Preparing informant versions is not simply a matter of changing "I" in self-report items to "He" or "She" in informant versions, although this works for some items, particularly those with high behavioral visibility (e.g., "I/He/She get(s) into more fights than most people"). Depending on the purpose of the informant assessment, items that reflect one's self-view (e.g., "I haven't made much of my life") may need to be phrased so that informants report on the target's self-view (e.g., "She thinks she hasn't made much of her life") or so that they report their own perspective about the person (e.g., "He hasn't made much of his life").

Similarly, informants can report on items that refer to internal experience (e.g., “I sometimes feel unreal”) only if the person has talked about those experiences, so such items must reflect this fact (“She says that she sometimes feels unreal”). It also is important to note that self-informant correlations are typically modest (.20-.30) to moderate (.40-.60) for a wide variety of reasons (see Achenbach et al., 2005; De Los Reyes et al., 2015; and Connelly & Ones, 2010 for meta-analyses and discussions).

Adaptations for different age groups.—Consistency of measurement across developmental periods requires measure adaptation and revalidation. Many such adaptations are extensions downward to adolescence of measures developed for adults (e.g., Linde, Stringer, Simms, & Clark, 2013). For children, measure adaptation and revalidation are enormously complex due to the need for (1) multi-source assessment—child self-report except in very young children, multiple informant reports (e.g., parent, teacher) that typically show low-to-moderate agreement (De Los Reyes et al., 2015), and behavioral observation—and (2) to address developmental changes in both the nature of the construct itself, and children’s levels on the construct of interest (e.g., Putnam et al., 2008). For the elderly, adaptations may or may not be needed to account for normal age-related changes. For example, if the research question is how does life satisfaction change across the life span, then adjustments for age-related change in physical health would confound the results. In contrast, when assessing psychopathology, normal-range age-related decline in physical ability needs to be considered lest it be wrongly interpreted as a psychological symptom. See Achenbach, Ivanova, and Rescorla (2017) for a research program on multicultural, multi-informant assessment of psychopathology across the lifespan.

Structural Validity: Item Selection and Psychometric Evaluation

Test-construction strategies.

Choosing a test-construction or item-selection strategy should match the scale development goal and the target construct(s)’ theoretical conceptualization. Loevinger (1957) described three main conceptual models: (a) quantitative (dimensional) models that differentiate individuals with respect to degree or level of the target construct, (b) class models that seek to categorize individuals into qualitatively different groups, and (c) more complex dynamic models. However, since then, including since the publication of our previous article, considerable research has confirmed that dimensional models fit the vast majority of data best (Haslam, Holland, & Kuppens, 2012; Markon, Chmielewski, & Miller, 2011a, 2011b), so anyone considering either a class or a dynamic model should have a very well-established theoretical reason to pursue it; consequently, we do not discuss them further in this article.

Loevinger (1957) championed the concept of structural validity (see also Messick, 1995)—that a scale’s internal structure (i.e., interitem correlations) should parallel the external structure of the target trait (i.e., correlations among non-test manifestations of the trait), and that items should reflect the underlying (latent) trait variance. These concerns parallel the three main item-selection strategies for dimensional constructs: empirical (primarily concerned with nontest manifestations), internal consistency (concerned with interitem structure), and item response theory (focused on latent traits). These methods are not mutually exclusive and typically should be used in conjunction with one another: structural validity encompasses all three.

Criterion-based methods.—Beginning with Meehl's (1945) "empirical manifesto," empirically keyed test construction became the dominant scale-construction method. Due to major difficulties in cross-validation and generalization, plus the method's inability to advance psychological theory, however, its popularity waned. The field readily embraced Cronbach and Meehl's (1955) notion of construct validity, although a number of years passed before the widespread availability of computers facilitated a broad switch to internal consistency methods of scale construction. Today, attention to empirical correlations with non-test criteria has largely shifted to the external validation phase of test development, although there is no reason to avoid examining these relations early in scale development. One common strategy is to administer the initial item pool to both community and clinical samples, and to consider differences in items' mean levels across the two groups as one criterion, among several, for item selection.

Internal-consistency methods.—Currently, the single most widely used method for item selection is some form of internal-consistency analysis. When developing a single scale, corrected item-total correlations are used frequently to eliminate items that do not correlate strongly with the assessed construct, but factor analysis is essential when the target construct is conceptualized as part of a hierarchical structure or when multiple constructs are being developed simultaneously. We typically use exploratory factor analysis (EFA) to identify the underlying latent dimension (for a single unidimensional scale) or dimensions (for a higher order scale with lower order facets). We then use the identified dimension(s) as the basis for scale creation (e.g., Clark & Watson, 1995; Simms & Watson, 2007).

Consistent with general guidelines in the broader factor-analytic literature (e.g., Comrey, 1988; Fabrigar, Wegener, MacCallum & Strahan, 1999; Russell, 2002), we recommend using principal factor analysis (PFA vs. principal components analysis; PCA) as the initial extraction method (but see Cortina, 1993, for arguments favoring PCA). It also is helpful to examine both orthogonal and oblique rotations (Watson, 2012). For most purposes, we recommend eliminating items (a) with primary loadings below .35-to-.40 (for broader scales; below .45-to-.50 for narrower scales) and (b) that have similar or stronger loadings on other factors, although these guidelines may need to be relaxed in some circumstances (e.g., clinical measures in which it may be important to include low base-rate items). Resulting scales can then be refined using CFA.

Factor analysis is a tool that can be used wisely or foolishly. Fortunately, the nature of factor analysis is such that blind adherence to a few simple rules typically will lead to developing a decent (though not likely optimal) scale. Even with factor analysis, there is no substitute for good theory and careful thought. For example, as noted earlier, internal consistency and breadth are countervailing, so simply retaining the strongest loading items may not yield a scale that best represents the target construct. That is, if the top-loading items are highly redundant with one another, including them all will increase internal consistency estimates but also may create an overly narrow scale that does not assess the construct optimally. This represents another illustration of the attenuation paradox we discussed previously.

Similarly, if items that reflect the theoretical core of the construct do not correlate strongly with it or with each other in preliminary analyses, it is not wise simply to eliminate them without considering why they did not behave as expected: Is the theory inadequate? Are the items poorly worded? Is the sample non-representative in some important way? Are the items' base rates too extreme? and Are there too few items representing the core construct?

Item Response Theory (IRT).—IRT (Reise et al., 2005; Reise & Waller, 2009; Simms & Watson, 2007) increasingly is being used in scale development; as noted earlier, it plays a particularly important role in short-form creation. IRT is based on the assumption that item responses reflect levels of an underlying construct and, moreover, that each item's response-trait relation can be described by a monotonically increasing function called an *item characteristic curve* (ICC). Individuals with higher levels of the trait have greater expected probabilities for answering the item in a keyed direction (e.g., highly extraverted individuals are more likely to endorse an item about frequent partying). ICCs provide the precise values (within a standard error range) of these probabilities across the entire range of trait levels.

In using IRT, the emphasis is on identifying the specific items that are maximally informative for each individual, given his or her level of the underlying dimension. For instance, a challenging algebra problem may provide useful information for respondents with a high level of mathematical ability (who may or may not be able to get it correct), but it is uninformative when given to individuals with little mathematical facility, because we know in advance that they almost surely will get it wrong. From an IRT perspective, the optimal item is one that a respondent has a 50% probability of answering correctly or endorsing in the keyed direction, because this provides the greatest increment in trait-relevant information for that person.

Within the IRT literature, a model with parameters for *item difficulty* and *item discrimination* is used most frequently (Reise & Waller, 2009; Simms & Watson, 2007). Item difficulty is the point along the underlying continuum at which an item has a 50% probability of being answered correctly (or endorsed in the keyed direction) across all respondents. Items with high (vs. low) difficulty values reflect higher (vs. lower) trait levels and have low (vs. high) correct-response/ endorsement probabilities. Discrimination reflects the degree of psychometric precision, or information, that an item provides across difficulty levels.

IRT offers two important advantages over other item-selection strategies. First, it enables specification of the trait level at which each item is maximally informative. This information can then be used to identify a set of items that yield precise, reliable, and valid assessment across the entire range of the trait. Thus, IRT methods offer an enhanced ability to discriminate among individuals at the extremes of trait distributions (e.g., both among those very high and very low in extraversion). Second, IRT methods allow estimation of each individual's trait level without administering a fixed set of items. This flexibility permits the development of *computer-adaptive tests* (CATs) in which assessment focuses primarily on the subset of items that are maximally informative for each respondent (e.g., difficult items for quantitatively gifted individuals, easier items for those low in mathematical ability). CATs are extremely efficient and provide roughly equivalent trait-relevant information using fewer items than conventional measures (typical item reductions are ≥ 50%; Reise & Waller, 2009; Rudick, Yam, & Simms, 2013).

As a scale development technique, IRT's main limitation is that it requires a good working knowledge of the basic underlying trait(s) to be modeled (i.e., that one's measurement model be reasonably well established). Consequently, IRT methods are most useful in domains in which the basic constructs are well understood and less helpful when the underlying structure is unclear. Thus, EFA remains the method of choice during the early, investigative stages of assessment within a domain. Once the basic factors/scales/constructs within the domain have been established, they can be refined further using analytic approaches such as CFA and IRT.

It also is more challenging to develop multiple scales simultaneously in IRT than using techniques such as EFA. Consequently, IRT-based scales frequently are developed in isolation from one another, with insufficient attention paid to discriminant validity. For example, Pilkonis et al. (2011) describe the development of brief, IRT-based anxiety and depression scales. Although these scales display some exemplary qualities and compare favorably in many ways to more traditional indicators of these constructs, the depression and anxiety trait scores correlated .81 with one another (Pilkonis et al., 2011). Similarly, they correlated .79 in a sample of 448 Mechanical Turk (MTurk) workers (Watson, Stanton, & Clark, 2017). This example illustrates the importance of developing scales simultaneously—using techniques such as EFA—to maximize measures' discriminant validity.

Initial data collection.

In this section, "initial" does not include pilot testing, which can be helpful to conduct on moderately sized samples of convenience (e.g., 100–200 college students or an MTurk sample) to test item formats, ensure that links to online surveys work, and so on.

Sample considerations.—Basic item-content decisions that will shape the scale's empirical and conceptual development are made after the first full round of data collection. Therefore, it is very important to use *at least* one, and *preferably* two or three, large, reasonably heterogeneous sample(s). Based on evidence regarding the stability and replicability of structural analyses (Guadagnoli & Velicer, 1988; MacCallum, Widaman, Zhang, & Hong, 1999), we recommend a minimum of 300 respondents per (sub)sample, including students, community adults (e.g., MTurk), and, ideally, a sample specifically representing the target population. If that is not feasible (which is not uncommon for financial reasons), then at least one sample should include individuals who share a key attribute with the target population. For example, if the measure is to be used in a clinical setting, then consider using an analog sample (e.g., college students or MTurk workers) who score above a given cut point on a psychopathology screening scale.

As we discuss below, the reason that it is critical to obtain such data early on is that the target construct and its items may have rather different properties in such samples compared to college-student or community-adult samples. If this is not discovered until late in the development process, the scale's utility may be seriously compromised.

Inclusion of comparison (anchor) scales.—In the initial round of data collection, it is common practice to administer the preliminary item pool without additional items or scales. This practice is regrettable, however, because it does not permit examination of the target construct's boundaries, which is critical to understanding the construct from both theoretical and empirical viewpoints. Just as the initial literature review permits identification of existing scales and concepts that may help establish the measure's convergent and discriminant validity, marker scales assessing these other constructs should be included in the initial data collection to begin to test these hypotheses. Too often, test developers belatedly discover that their new scale correlates too strongly with an established measure or, worse, with one of a theoretically distinct construct.

Scale co-development.—As our knowledge of the hierarchical structure of personality and psychopathology has grown, paying attention to where a construct fits within a particular structure has become important enough that in most cases, new scales should not be developed in isolation, but rather "co-developed" with the express intent of considering their convergent and discriminant validity as part of the initial scale-development process. Note that single scales with subscales are special examples of hierarchical structures, so the various principles we discuss here are relevant to such cases as well; we specifically address subscales subsequently.

When co-developing scales for constructs conceptualized as fitting within a hierarchical structure, there are several issues to consider initially, the most important of which are (a) whether the primary focus of scale development is on the lower order constructs, the higher order constructs, or both; and (b) how extensive a portion of the hierarchical structure the scale-development project targets. For example, in developing the SNAP (Clark et al., 2014), the focus was on lower order scales relevant to personality pathology which, accordingly, were developed without regard to how they loaded on higher order dimensions. In contrast, in developing the Faceted Inventory of the Five-Factor Model (FI-FFM; Watson Nus, & Wu, 2019), the focus was equally on the higher and lower order dimensions, so scales that either cross-loaded on two or more higher order dimensions or did not load strongly on any higher order dimension were not included in the final instrument.[4]

In developing the Personality Inventory for *DSM-5* (PID-5; Krueger, Derringer, Markon, Watson, & Skodol, 2012), the facets of the five hypothesized domains were initially created individually. Then, exploratory factor analyses (EFA) were run on all items within each domain, extracting factors up to one more than the number of each domain's hypothesized facets, and selecting the EFA with the best fit to the data for each domain. In several cases, several facets were collapsed into one scale because their respective items formed a single factor. In other cases, items were moved from one facet to another. Less than half the facets "survived" these analyses more-or-less intact; in the most extreme case, five facets merged into one.

[4] Note that the former and latter approaches facilitate and eschew, respectively, including orphan and interstitial constructs, which we discussed earlier.

However, even this level of attention proved to be insufficient. A scale-level EFA with oblique rotation found significant cross-factor loadings for 11 (44%) of the PID-5's 25 facets, and subsequent studies have established that the measure would have benefitted from additional examination of discriminant validity across domains. Crego, Gore, Rojas, and Widiger (2015) provided a brief review of such studies and reported that in their own data, five PID-5 facets (20%) had higher average discriminant (cross-domain) than convergent (within-domain) correlations. In the Improving the Measurement of Personality Project (IMPP), with a mixed sample of 305 outpatients and 302 community adults screened to be at high risk for personality pathology, we replicated those results for four of these five facets (Clark, 2018). Of course, it may be that, although conceptualized initially as facets of a particular domain, these traits actually are interstitial and thus should have high cross-loadings. Our point here is not to set the bar for scale development prohibitively high, but to raise attention to important issues such as paying careful attention to expected and observed convergent and discriminant validity, so as to maximize the outcomes of scale development projects and increase the likelihood that new scales will make important contributions to the psychological literature.

Subscales.—A special case of co-development is that of *hierarchically multidimensional* scales, commonly called measures with subscales. These measures are unidimensional at their higher order level, with correlated subfactors at their lower order level. Factor analysis of the items comprising a hierarchically multidimensional construct yields a strong general factor, with all or almost all items loading strongly on this factor; nonetheless, when additional factors are extracted, clear—yet correlated—lower order dimensions emerge. For example, two subscales constitute the SNAP-2 Self-harm scale—Suicide Potential and Low Self-esteem (Clark et al., 2014)—the former composed of item content directly concerned with suicide and self-harming thoughts, feelings, and behaviors, and the latter composed of items expressing self-derogation versus self-satisfaction. In the IMPP sample, all 16 items loaded most strongly on the first general factor (*M* loading = .52, range = .41 to .66), yet when two varimax-rotated factors were extracted, the two subscales' items loaded cleanly on the two factors, with mean loadings of .57 and .17 on their primary and non-primary factors, respectively (Clark, 2017); unit-weighted subscales constructed from the two subsets of items correlated .48. Any hierarchically dimensional scale (i.e., any measure with subscales) should have similar properties.

Psychometric evaluation: An iterative process.

We return here to an earlier point: Good scale construction is an iterative process involving an initial cycle of preliminary measure development, data collection, and psychometric evaluation, followed by at least one additional cycle of revision of *both* measure *and* construct, data collection, psychometric evaluation, revision. . .. The most often neglected aspect of this process is revision of the target construct's conceptualization. Too often, scale developers assume that their initial conceptualization is entirely correct, considering only the measure as open to revision. However, it is critical to remain open to rethinking one's initial construct—to "listen to the data" not "make the data talk."

Often this involves only slight tweaking, but it may involve more fundamental reconceptualization. For example, the Multidimensional Personality Questionnaire (Tellegen & Waller, 2008) originally had a single, bipolar, trait-affect scale, but ended with nearly orthogonal negative and positive emotionality scales. The SNAP-2 Self-harm scale provides a converse example. Initially two scales, Low Self-esteem (originally Self-derogation) and Suicide Proneness were developed independently. However, across repeated rounds of data collection in diverse samples, the scales correlated highly and yielded a single dimension with two lower order facets, so they were combined to form a Self-harm scale, with two subscales. This necessitated reconceptualization of the combined item set as single construct with two strongly correlated item-content subsets. Research on self-injurers provided an initial basis for this reconceptualization: Glenn, Michel, Franklin, Hooley, and Nock (2014) found that the relation between self-injury and experimentally tested pain analgesia was mediated by self-criticalness and hypothesized “the tendency to experience self-critical thoughts in response to stressful events . . . increases the likelihood of both self-injury and pain analgesia” (p. 921). A full reconceptualization lies in the future, but the convergence of findings from independently conducted self-injury research and psychometric scale evaluation is evidence supporting both.

Analysis of item distributions.—Before conducting more complex structural analyses, scale developers should examine individual items’ response distributions. Two considerations are paramount: First, it is important to consider eliminating items that have highly skewed and unbalanced distributions. In a true/false format, these are items that virtually everyone (e.g., 95% or more) either endorses or denies; with a Likert-rating format, these are items to which almost all respondents respond similarly (e.g., “slightly agree”). Highly unbalanced items are undesirable for several reasons: (1) When most respondents answer similarly, items convey very little information, except perhaps at extremes of the trait distribution; (2) relatedly, items with limited variability are likely to correlate weakly with other items, and therefore will fare poorly in structural analyses; and (3) items with extremely unbalanced distributions can produce highly unstable correlational results (see Clark & Watson, 1995, for an example from Comrey, 1988).

Importantly, only items with unbalanced distributions across diverse samples representing the full range of the scale’s target population should be eliminated. As mentioned earlier, many items show very different response distributions across clinical and nonclinical samples. For instance, the item “I have things in my possession that I can’t explain how I got” likely would be endorsed by very few undergraduates, whereas in an appropriate patient sample, it may have a much higher endorsement rate and prove useful in assessing clinically significant levels of dissociative pathology. Thus, it may be desirable to retain items that assess important construct-relevant information in a sample more like the target population, even if they have extremely unbalanced distributions and, therefore, relatively poor psychometric properties in others.

The second consideration is that it is desirable to retain items with a broad range of distributions. In the case of true/false and Likert-type items, respectively, this means keeping items with widely varying endorsement percentages and means (in IRT terms, items with widely varying difficulty parameters), because most constructs represent continuously distributed dimensions, such that scores can occur across the entire dimension. Thus, it is important to retain items that discriminate at many different points along the continuum (e.g., at mild, moderate, and extreme levels). Returning to the earlier example, “I have things in my possession that I can’t explain how I got” may be useful precisely because it serves to define the extreme upper-end of the dissociative continuum (i.e., those who suffer from dissociative identity disorder).

As noted earlier, a key advantage of IRT (Reise et al., 2005; Reise & Waller, 2009) is that it yields parameter estimates that specify the point along a continuum at which a given item is maximally informative. These estimates can be used to choose an efficient set of items that yield precise assessment across the entire range of the continuum, which naturally leads to retaining items with widely varying distributions.

Unidimensionality, internal consistency, and coefficient alpha.—The next stage is to determine which items to eliminate or retain in the item pool via structural analyses. This is most critical when seeking to create a theoretically based measure of a target construct, so that the goal is to measure one thing (i.e., the target construct)—and *only* this thing—as precisely as possible. This goal may seem relatively straightforward, but it remains poorly understood by test developers and users. The most obvious problem is the widespread misapprehension that this goal can be attained simply by demonstrating an “acceptable” level of internal consistency reliability, typically as estimated by coefficient alpha (Cronbach, 1951). A further complication is that recommendations regarding .80 as “acceptable” level of internal consistency for basic research (e.g., Nunnally, 1978; Streiner, 2003) are widely ignored, such that characterizations of coefficient alphas in the .60s and .70s as “good” or “adequate” are far too common.

More fundamentally, psychometricians long have disavowed using reliability indexes to establish scales’ homogeneity (see Boyle, 1991; Cortina, 1993). To understand why this is so, we must distinguish between *internal consistency* and *homogeneity* or *unidimensionality*. “Internal consistency” refers to the overall degree to which a scale’s items are intercorrelated, whereas “homogeneity” and “unidimensionality” indicate whether or not the scale items assess a single underlying factor or construct (Briggs & Cheek, 1986; Cortina, 1993). Thus, internal consistency is a necessary but not sufficient condition for homogeneity or unidimensionality. In other words, a scale cannot be homogeneous unless all of its items are interrelated. Because theory-driven assessment seeks to measure a single construct systematically, the test developer ultimately is pursuing the goal of homogeneity or unidimensionality, not internal consistency per se.

Unfortunately, KR-20 and coefficient alpha are measures of internal consistency, not homogeneity, and so are of limited utility in establishing the unidimensionality of a scale. Furthermore, they are even ambiguous and imperfect indicators of internal consistency, because they essentially are a function of two parameters: scale length and the average interitem correlation (AIC; Cortina, 1993; Cronbach, 1951). Thus, one can achieve a high internal consistency reliability estimate with many moderately correlated items, a small number of highly intercorrelated items, or various combinations of scale length and AIC. Whereas AIC is a straightforward indicator of internal consistency, scale length is entirely irrelevant. In fact, with a large number of items, it is exceedingly difficult to *avoid* having a high reliability estimate, so coefficient alpha is virtually useless for scales containing 40 or more items (Cortina, 1993).

Accordingly, the AIC is a much more useful index than coefficient alpha, and test developers should work toward a target AIC, rather than a particular level of alpha. As a more specific guideline, we recommend that the AIC fall in the range of .15 to .50 (see Briggs & Cheek, 1986), with the scale's optimal value determined by the generality versus specificity of the target construct. For a broad higher order construct such as extraversion, a mean correlation as low as .15-.20 may be desirable; by contrast, for a valid measure of a narrower construct such as talkativeness, a much higher mean intercorrelation (e.g., in the .40-to-.50 range) is needed.

As suggested earlier, however, even the AIC cannot alone establish the unidimensionality of a scale; in fact, a multidimensional scale actually can have an "acceptable" AIC: Cortina (1993, Table 2) artificially constructed an 18-item scale composed of two distinct 9-item groups. The items within each cluster had an AIC of .50. However, the two clusters were uncorrelated. Obviously, the full scale was not unidimensional, instead reflecting two completely independent dimensions; nevertheless, it had a coefficient alpha of .85 and a moderate AIC (.24).

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

Cross-validation.—If our previous recommendations have been followed, this section would hardly be necessary, because we advise testing the item pool and resultant scales in multiple samples from essentially the beginning of the process. Cross-validation has been made much easier by the existence of crowdsourcing platforms such as MTurk, and it appears that most scale-development articles published these days in top-line journals such as *Psychological Assessment*, include a cross-validation sample. However, we easily were able to identify some that did not, even among articles new enough to be in the "Online First" publication stage.

Loevinger split the structural and external stages at the point wherein the focus moves from items to total scores. We follow her lead and shift now to the external phase of development.

External Validity: An Ongoing Process

We have emphasized the iterative process of scale development. Phrased in its most extreme form, scale development ends only when a measure is “retired” because, due to increased knowledge, it is better to develop a new measure of a revised construct than to modify an existing measure. That said, when it has been established that measures have strong psychometric properties relative to their respective target constructs, evaluation shifts to focusing on placement in their immediate and broader nomological net (Cronbach & Meehl, 1955). As our focus is primarily on scale development, we cover only a few important aspects of this stage. We refer readers to Smith and McCarthy (1995), who describe the later “refinement” stages of scale development in some detail.

First, however, it is important to note that the quality of the initial scale-development stages has clear ramifications for external validity. If the concept is clearly conceptualized and delineated, its “rival” constructs and target criteria will also be clearer. If the original item pool included a widely relevant range of content, the scale’s range of clinical utility will be more clearly defined. If the measure was constructed with a focus on unidimensionality (vs. internal consistency), the scale will identify a more homogeneous clinical group, rather than a more heterogeneous one requiring further demarcation. Finally, if convergent and discriminant validity have been considered from the outset, it will be far easier to delineate the construct boundaries and achieve the important goal of knowing exactly what the scale measures and what it does not.

Convergent and discriminant validity.

According to the *Test Standards,* “Relationships between test scores and other measures intended to assess the same or similar constructs provide convergent evidence, whereas relationships between measures purportedly of different constructs provide discriminant evidence” (AERA, APA, & NCME, 2014, pp. 16–17). Inclusion of “similar constructs” in this definition creates an unfortunate gray area in which it is unclear whether a construct is similar enough to provide convergent—as opposed to discriminant—evidence. It is clearer simply to state that convergent validity is assessed by examining relations among purported indicators of the same construct. Using Campbell and Fiske’s (1959) terminology, convergent evidence is established by examining monotrait correlations.

Campbell and Fiske (1959) argue that convergent correlations “should be significantly different from zero and sufficiently large to encourage further examination of validity” (p. 82). Unfortunately, what “sufficiently large” means in this context cannot be answered simply because the expected magnitude of these correlations will vary dramatically as a function of various design features. The single most important factor is the nature of the different methods that are used to examine convergent validity. In their original formulation, Campbell and Fiske (1959) largely assumed that investigators would examine convergence across fundamentally different methods. For example, in one analysis (their Table 2), they examined the associations between trait scores assessed using (a) peer ratings versus (b) a word association task.

Over time, investigators began to interpret the concept of “method” much more loosely, for example, offering correlations among different self-report measures of the same target construct to establish convergent validity (e.g., Watson et al., 2019; Watson et al., 2007). This practice is not problematic, but obviously is very different from what Campbell and Fiske (1959) envisioned. Most notably, convergent correlations will be—and should be—substantially higher when they are computed within the same basic method (e.g., between different self-report measures of neuroticism) than when they are calculated across very different methods (e.g., between self- vs. informant-rated neuroticism). This, in turn, means that the same level of convergence might support construct validity in one context, but challenge it in another. For instance, it would be difficult to argue that a .45 correlation between two self-report measures of self-esteem reflects adequate convergent validity, but the same correlation between self- and parent-ratings of self-esteem might do so.

Discriminant validity involves examining how a measure relates to purported indicators of other constructs (i.e., heterotrait correlations). Discriminant validity is particularly important in establishing that highly correlated constructs within hierarchical models are, in fact, empirically distinct from one another (see, for example, Watson & Clark, 1992; Watson et al., 2019). Indeed, the most interesting tests of discriminant validity involve near-neighbor constructs that are known to be strongly related to one another (Watson, 2012).

Campbell and Fiske (1959) state that discriminant validity is established by demonstrating that convergent correlations are higher than discriminant coefficients. For instance, self-rated self-esteem should correlate more strongly with peer-rated self-esteem than with peer-rated extraversion. One complication, however, is that the meaning of the word “higher” is ambiguous in this context. Many researchers interpret it rather loosely to mean simply that the convergent correlation must be *descriptively higher* than the discriminant correlations to which it is compared. For instance, if the convergent correlation is .50, and the highest discriminant correlation is only .45, then it is assumed that this requirement is met.

It is better to use the more stringent requirement that the convergent correlation should be *significantly higher* than the discriminant coefficients to which it is compared, which obviously is more difficult to meet. Perhaps most importantly, it also requires relatively large sample sizes (typically, at least 200 observations) to have sufficient statistical power to conduct these tests in a meaningful way. Nevertheless, the payoff is well worth it in terms of the greater precision of the validity analyses. For instance, Watson et al. (2008) examined the convergent and discriminant validity of the 11 non-overlapping IDAS scales in a sample of 605 outpatients. The convergent correlations ranged from .52 to .71, with a mean value of .62. Significance tests further revealed that these convergent correlations exceeded the discriminant coefficients in 219 of 220 comparisons (99.5%). These results thereby provide substantial evidence of discriminant validity.

Criterion validity.

Criterion validity is established by demonstrating that a test is significantly related to theoretically relevant non-test outcomes (e.g., clinical diagnoses, arrest records). Although criterion keying is no longer widely used as a scale-development method, demonstrating criterion validity remains an important part of construct validation. The choice of criteria represents a very important aspect of examining criterion validity. Specifically, it is important to put the construct to what Meehl (1978) called a “risky test” (p. 818), one that provides strong support for the construct if it passes. It is not uncommon for developers of a measure of some aspect of psychopathology to claim criterion validity based on finding significant differences between scores on the measure in target-patient and non-patient samples. This is not a risky test; for example, it would be far better to show that the measure differentiated a particular target-patient group from other types of patients.

Incremental validity.

Criterion validity also involves the related concept of incremental validity. Incremental validity is established by demonstrating that a measure adds significantly to the prediction of a criterion over and above what can be predicted by other sources of data (Hunsley & Meyer, 2003). Ensuring that a scale is sufficiently distinct from well-established constructs that it has significant incremental validity for predicting important external variables is a crucial issue that often is given too little consideration.

Three interrelated issues are important when considering incremental validity. They also are related to discriminant validity, so we discuss them together. The first issue is what variables to use to test incremental validity—most notably, what are its competing predictors, and also what criterion is being predicted. The second is intertwined with where a measure fits within an established hierarchical structure; a full understanding of a new measure’s incremental validity requires comparison with other measures at the same level of abstraction. The third is how much incremental validity is enough, which affects interpretation of findings and the new measure’s value to the field. When considering incremental and discriminant validity, it again is important to put the construct to a “risky test” (Meehl, 1978). For incremental validity, this means adding significant predictive power to a known, strong correlate of the criterion, whereas for discriminant validity, this means showing independence from variables that one might expect would be correlated with the new measure. The biggest challenge, therefore, is to demonstrate discriminant *and* incremental validity for a new measure of an established construct.

We again use *grit* (Duckworth et al., 2007) to illustrate. In both its seminal article and that introducing its short form (Duckworth & Quinn, 2009), grit’s developers acknowledged that conscientiousness (C) was highly correlated with grit (r = ~.70-.75). Conscientiousness was thus a strongly competing predictor, raising concerns about both incremental and discriminant validity; nonetheless, grit showed incremental validity over C in two studies of relevant criteria. However, although Duckworth and Quinn (2009) acknowledge that grit might not outpredict C’s facets, they did not examine this question empirically. Subsequently, MacCann and Roberts (2010) reported that grit had no incremental validity over eight C facets for predicting a number of relevant variables. Moreover, meta-analytic results (Credé, Tynan, & Harms, 2017) indicated that grit had insufficient discriminant validity and “was simply a different manifestation of conscientiousness” (p. 12). These results reinforce the importance of considering which level of an established hierarchy provides the “riskier” test of incremental validity; such tests also serve to determine where a new construct fits best within that hierarchy.

Watson et al. (2019) provide a good recent example of “risky” tests of incremental validity involving the FI-FFM neuroticism facet scales. The authors first presented convergent and discriminant validity analyses to establish that some scales assessed the same trait dimensions as scales in the NEO Personality Inventory-3 (NEO-PI-3; McCrae, Costa, & Martin, 2005), whereas others did not. They then reported incremental-validity analyses wherein they showed that the two FI-FFM scales assessing novel traits—Somatic Complaints and Envy—added significantly to the prediction of various criteria (e.g., anxiety and depressive disorder diagnoses, measures of health anxiety and hypochondriasis) over and above the NEO-PI-3 facets.

Finally, the issue of whether a construct has sufficient incremental validity for predicting relevant constructs has no simple answer. Rather, it depends on the purpose and scope of prediction and ultimately reduces to a type of cost-benefit analysis. For example, in epidemiological studies, relatively weak predictors with small but significant incremental validity (e.g., 2–3%) may lead to public-health-policy recommendations that, if followed, could save thousands of lives. Thus, in medicine, statistics such as “number needed to treat” (NNT; a prediction of the number of additional people who would need to be treated to affect one of them) have been developed. But whether 10 or 100 is a small enough number to warrant treating more people depends on the cost of the treatment, the amount of harm caused by not treating, and the degree of benefit to those treated successfully. For a low-cost treatment with great harm for not treating and great benefit for successful treatment (e.g., a cheap vaccine for a usually fatal illness), a very large NNT might still be small enough. Conversely, for a very expensive treatment with modest harm and benefit (e.g., an experimental treatment that only slightly prolongs life), respectively, a very small NNT would be more appropriate. Use of such statistics in psychological research would help to clarify evaluating new measures’ incremental validity.

Cross-method analyses.

The utility of obtaining information from multiple sources is increasingly being recognized, and research related to the issues that arise when diverse sources provide discrepant information also is growing, but the topic still remains relatively understudied. Earlier, we discussed informant reports, and we do not have much additional space to devote to the broader topic of cross-method analyses, so we simply bring a few issues to readers’ attention.

Self-report questionnaires versus interviews.—Self-ratings often are denigrated unfairly as “just self-report,” whereas, for many psychological phenomena, self-report is the best—or even the only—appropriate method. For example, no one knows how a person is feeling or how much something hurts other than that person, but inherent in this strength are self-report’s greatest limitations: Because N always = 1 for information regarding an individual’s internal sensations, no other method is available to verify such self-reports, even though (1) individuals may not report on their internal thoughts or feelings accurately, either because they choose not to do so (e.g., if honest reporting might lead to an adverse decision) or because they cannot (e.g., poor insight, memory lapses, dementia); (2) individuals may interpret and use rating scales differently (e.g., “usually” may represent different subjective frequencies across individuals; Schmidt, Le, & Ilies, 2003); and (3) individual items may be interpreted differently depending on one's level of the trait. For example, the item "I'm often not as cautious as I should be" was intended as an indicator of impulsivity, but was endorsed positively more often by respondents *lower* on the trait (i.e., more cautious individuals!; Clark et al., 2014). Nonetheless, despite these limitations, self-reported information is remarkably reliable and supports valid inference most of the time.

Interviews often are considered superior to self-report because (a) they involve "expert judgment" and (b) follow-up questions permit clarification of responses. However, for the most part, they are based on self-report and thus reflect the strengths and limitations of both self-report and interviewing. Perhaps the most serious limitation of interviews is that interviewers always filter interviewees' responses through their own perspective and, as the clinical-judgment literature has shown repeatedly (e.g., Dawes, Faust, & Meehl, 2002), this typically lowers reliability and predictive validity over an empirically established method, a problem that decreases relative to the degree of structure in the interview. For example, the *DSM-III* field trials reported unstructured-interview-based interrater and retest reliabilities for personality disorder of .61 and .54, respectively (Spitzer, Forman, & Nee, 1979), whereas a later review based on semi-structured interviews reported these values to be .79 and .62 (Zimmerman, 1994). Convergence with self-report also drops when unstructured versus semi-structured interviews are compared; for example, Clark, Livesley, and Morey (1997) reported mean correlations of .25 versus .40 and mean kappas of .08 versus .27 for these comparisons. Nonetheless, demonstrating good convergent/ discriminant validity between self-report and interview measures of constructs provides support for both methods (e.g., Dornbach-Bender et al., 2017; Watson et al., 2008).

Symptom scales versus diagnoses.—There often is a parallelism of symptom and diagnoses with self-report versus interview that is important to consider in its own right. However, we focus here only on interview-based assessment of both. Because they are more fully dimensional than dichotomous diagnoses, symptom scales are both more reliable and valid (Markon et al., 2011a, b). However, diagnoses carry a greater degree of clinical "respectability," which we believe is unwarranted, but is a reality nonetheless. One important issue is the use of "skip-outs": not assessing a symptom set if a core criterion is not met. To the extent possible, we recommend against using skip-outs because there often is important information in symptoms even when a core criterion is not met (Dornbach-Bender et al., 2017; Kotov, Perlman, Gámez, & Watson, 2015), with some clear exceptions, such as trauma-related symptoms. Demonstrating that similar inferences can be made from symptom measures and diagnoses increases the credibility of the symptom measures, so we recommend comparing them when feasible, all the while recognizing that the obtained correlations are simply indicators of convergent validity between two measures of the same phenomenon, not a comparison of a proxy against a gold standard.

Conclusion

We concluded Clark and Watson (1995) by noting that both the target of measurement and measurement of the target are important for optimal scale development, that later stages will proceed more smoothly if the earlier stages have both theoretical clarity and empirical precision. These points are still important today, but we now also encourage scale developers to consider the broader context of their target construct, in particular, the reasonably well-established hierarchical structures of personality and, to a lesser but growing extent, psychopathology.

Due to expansion of knowledge in our field, it is increasingly important to attend to measures' external validity, particularly convergent and discriminant validity, and incremental validity over well-established measures; as well as to use multi-trait, multi-method, multi-occasion frameworks for evaluating new measures. Perhaps we can summarize the direction in which scale development is moving by stating that in the fields of personality and psychopathology, the nomological net is no longer just an abstract ideal to which we need only pay lip service, but a practical reality that deserves our full and careful consideration.

Acknowledgments

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

The study reported herein was supported by National Institute of Mental Health Grant R01-MH068472 to David Watson.

References

Achenbach TM, Ivanova MY, & Rescorla LA (2017). Empirically based assessment and taxonomy of psychopathology for ages 1½–90+ years: Developmental, multi-informant, and multicultural findings. Comprehensive Psychiatry, 79, 4–18. 10.1016/j.comppsych.2017.03.006 [PubMed: 28356192]

Achenbach TM, Krukowski RA, Dumenci L, & Ivanova MY (2005). Assessment of adult psychopathology: Meta-analyses and implications of cross-informant correlations. Psychological Bulletin, 131(3), 361–382. doi: 10.1037/0033-2909.131.3.361 [PubMed: 15869333]

Ahmed SR, Fowler PJ, & Toro PA (2011). Family, public and private religiousness and psychological well-being over time in at-risk adolescents. Mental Health, Religion & Culture, 14(4), 393–408. doi: 10.1080/13674671003762685

Allport GW, & Odbert HS (1936). Trait-names: A psycho-lexical study. Psychological Monographs, 47(1), i–171. doi: 10.1037/h0093360

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (AERA, APA, & NCME). (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

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

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

Boyle GJ (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? Personality and Individual Differences, 12(3), 291–294.doi: 10.1016/0191-8869(91)90115-R

Briggs SR, & Cheek JM (1986). The role of factor analysis in the development and evaluation of personality scales. Journal of Personality, 54, 106–148.> doi: 10.1111/j.1467-6494.1986.tb00391

Brown A, & Maydeu-Olivares A (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. Psychological Methods, 18(1), 36–52. doi: 10.1037/a0030641 [PubMed: 23148475]

Burisch M (1984). Approaches to personality inventory construction: A comparison of merits. American Psychologist, 39, 214–227. doi: 10.1037/0003-066X.39.3.214

Campbell DT, & Fiske DW (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56(2), 81–105. 10.1037/h0046016 [PubMed: 13634291]

Carmona-Perera M, Caracuel A, Pérez-García M, & Verdejo-García A (2015). Brief moral decision-making questionnaire: A Rasch-derived short form of the Greene dilemmas. Psychological Assessment, 27(2), 424–432. doi: 10.1037/pas0000049 [PubMed: 25558971]

Clark LA, Livesley WJ, & Morey L (1997). Personality disorder assessment: The challenge of construct validity. Journal of Personality Disorders, 11, 205–231. 10.1521/pedi.1997.11.3.205 [PubMed: 9348486]

Clark LA (2018). The Improving the Measurement of Personality Project (IMPP). Unpublished dataset. University of Notre Dame, Notre Dame, IN.

Clark LA, Simms LJ, Wu Kevin, D., & Casillas A (2014). Schedule for Nonadaptive and Adaptive Personality-2nd Edition (SNAP-2): Manual for Administration, Scoring, and Interpretation. Notre Dame, IN: University of Notre Dame.

Clark LA, & Watson DB (1995). Constructing validity: Basic issues in objective scale development. Psychological Assessment, 7, 309–319. doi: 10.1037/1040-3590.7.3.309

Comrey AL (1988). Factor-analytic methods of scale development in personality and clinical psychology. Journal of Consulting and Clinical Psychology, 56, 754–761. doi: 10.1037/0022-006X.56.5.754 [PubMed: 3057010]

Connelly BS, & Ones DS (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. Psychological Bulletin, 136(6), 1092–1122. doi: 10.1037/a0021212 [PubMed: 21038940]

Costa PT, & McCrae RR (1992). Revised NEO Personality Inventory (NEO-PIR) and NEO Five Factor Inventory (NEO-FFI) professional manual. Odessa, FL: Psychological Assessment Resources.

Cortina JM (1993). What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology, 78, 98–104. doi: 10.1037/0021-9010.78.1.98

Credé M, Tynan MC, & Harms PD (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. Journal of Personality and Social Psychology, 113(3), 492–511. 10.1037/pspp0000102 [PubMed: 27845531]

Cronbach LJ (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297–334. doi: 10.1007/BF02310555

Cronbach LJ, & Meehl PE. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281–302. 10.1037/h0040957 [PubMed: 13245896]

Dawes RM, Faust D, & Meehl PE (2002). Clinical versus actuarial judgment In Gilovich T, Griffin D & Kahneman D (Eds.), Heuristics and biases: The psychology of intuitive judgment; heuristics and biases: The psychology of intuitive judgment (pp. 716–729, Chapter xvi, 857 Pages) Cambridge University Press, New York, NY.

De Los Reyes A, Augenstein TM, Wang M, Thomas SA, Drabick DAG, Burgers DE, & Rabinowitz J (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. Psychological Bulletin, 141(4), 858–900. doi: 10.1037/a0038498 [PubMed: 25915035]

Dornbach-Bender A, Ruggero CJ, Waszczuk MA, Gamez W, Watson D, & Kotov R (2017). Mapping emotional disorders at the finest level: Convergent validity and joint structure based on alternative measures. Comprehensive Psychiatry, 79, 3139. 10.1016/j.comppsych.2017.06.011 [PubMed: 28754505]

Duckworth AL, Peterson C, Matthews MD, & Kelly DR (2007). Grit: Perseverance and passion for long-term goals. Journal of Personality and Social Psychology, 92(6), 1087–1101. doi: 10.1037/0022-3514.92.6.1087 [PubMed: 17547490]

Duckworth AL, & Quinn PD (2009). Development and validation of the short grit scale (GRIT–S). Journal of Personality Assessment, 91(2), 166–174. doi: 10.1080/00223890802634290 [PubMed: 19205937]

Fabrigar LR, Wegener DT, MacCallum RC, & Strahan EJ (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4(3), 272–299. doi: 10.1037/1082-989X.4.3.272

Finch JF, & West SG (1997). The investigation of personality structure: Statistical models. Journal of Research in Personality, 31(4), 439–485. doi: 10.1006/jrpe.1997.2194

Funder DC (2012). Accurate personality judgment. Current Directions in Psychological Science, 21(3), 177–182. doi: 10.1177/0963721412445309

Geisinger KF (2003). Testing and assessment in cross-cultural psychology. In Graham JR, & Naglieri JA (Eds.), Handbook of psychology: Assessment psychology, vol. 10 (pp. 95–117, Chapter xix, 630 Pages) John Wiley & Sons Inc., Hoboken, NJ.

Glenn JJ, Michel BD, Franklin JC, Hooley JM, & Nock MK (2014). Pain analgesia among adolescent self-injurers. Psychiatry Research, 220(3), 921–926. doi: 10.1016/j.psychres.2014.08.016 [PubMed: 25172611]

Green BF Jr. (1978). In defense of measurement. American Psychologist, 33, 664–670.doi: 10.1037/0003-066X.33.7.664

Green DP, Goldman SL, & Salovey P (1993). Measurement error masks bipolarity in affect ratings. Journal of Personality and Social Psychology, 64(6), 1029–1041. doi: 10.1037/0022-3514.64.6.1029 [PubMed: 8326466]

Guadagnoli E, & Velicer WF (1988). Relation to sample size to the stability of component patterns. Psychological Bulletin, 103(2), 265–275. doi: 10.1037/0033-2909.103.2.265 [PubMed: 3363047]

Haslam N, Holland E, & Kuppens P (2012). Categories versus dimensions in personality and psychopathology: A quantitative review of taxometric research. Psychological Medicine, 42, 903–920. doi: 10.1017/S0033291711001966 [PubMed: 21939592]

Haynes SN, Richard DCS, & Kubany ES. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. Psychological Assessment, 7(3), 238–247. 10.1037/1040-3590.7.3.238

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

Hu L, & Bentler PM (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6(1), 1–55. doi: 10.1080/10705519909540118

Hu L, & Bentler PM (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychological Methods, 3(4), 424–453. doi: 10.1037/1082-989X.3.4.424

Hunsley J, & Meyer GJ (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. Psychological Assessment, 15, 446–455. doi: 10.1037/1040-3590.15.4.446 [PubMed: 14692841]

Kotov R, Perlman G, Gamez W, & Watson D (2015). The structure and short-term stability of the emotional disorders: A dimensional approach. Psychological Medicine, 45, 1687–1698. doi: 10.1017/S0033291714002815 [PubMed: 25499142]

Krueger RF, Derringer J, Markon KE, Watson D, & Skodol AE. (2012). Initial construction of a maladaptive personality trait model and inventory for DSM-5. Psychological Medicine, 42(9), 1879–1890. 10.1017/S0033291711002674 [PubMed: 22153017]

Lee K, Ogunfowora B, & Ashton MC (2005). Personality traits beyond the big five: Are they within the HEXACO space? Journal of Personality, 73(5), 1437–1463.doi: 10.1111/j.1467-6494.2005.00354.x [PubMed: 16138878]

Linde JA, Stringer DM, Simms LJ, & Clark LA (2013). The Schedule for Nonadaptive and Adaptive Personality Youth version (SNAP-Y): Psychometric properties and initial validation. Assessment, 20(4), 387–404. doi: 10.1177/1073191113489847 [PubMed: 23794180]

Loevinger J (1954). The attenuation paradox in test theory. Psychological Bulletin, 51(5), 493–504. doi: 10.1037/h0058543 [PubMed: 13204488]

Loevinger J (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3, 635–694. doi: 10.2466/PR0.3.7.635-694

Lowe JR, Edmundson M, & Widiger TA (2009). Assessment of dependency, agreeableness, and their relationship. Psychological Assessment, 21(4), 543–553. doi: 10.1037/a0016899 [PubMed: 19947788]

MacCallum RC, Widaman KF, Zhang S, & Hong S (1999). Sample size in factor analysis. Psychological Methods, 4(1), 84–99. doi: 10.1037/1082-989X.4.1.84

MacCann C, & Roberts RD (2010). Do time management, grit, and self-control relate to academic achievement independently of conscientiousness? In Hicks R (Ed.), Personality and individual differences: Current directions (pp. 79–90). Queensland, Australia: Australian Academic Press.

Mansolf M, & Reise SP (2016). Exploratory bifactor analysis: The Schmid-Leiman orthogonalization and Jennrich-Bentler analytic rotations. Multivariate Behavioral Research, 51(5), 698–717. doi: 10.1080/00273171.2016.1215898 [PubMed: 27612521]

Markon KE, Chmielewski M, & Miller CJ (2011a). The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. Psychological Bulletin, 137(5), 856–879. doi: 10.1037/a0023678 [PubMed: 21574681]

Markon KE, Chmielewski M, & Miller CJ (2011b). “The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review”: Correction to Markon et al. (2011). Psychological Bulletin, 137(6), 1–1093. doi: 10.1037/a0025727 [PubMed: 21219055]

McCrae RR (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. Personality and Social Psychology Review, 19(2), 97–112.doi: 10.1177/1088868314541857 [PubMed: 24989047]

McCrae RR, Costa PT Jr., & Martin TA (2005). The NEO-PI-3: A more readable revised NEO personality inventory. Journal of Personality Assessment, 84(3), 261–270.doi: 10.1207/s15327752jpa8403_05 [PubMed: 15907162]

McDade-Montez E, Watson D, O’Hara MW, & Denburg NL (2008). The effect of symptom visibility on informant reporting. Psychology and Aging, 23(4), 940–946.doi: 10.1037/a0014297 [PubMed: 19140663]

Meehl PE (1945). The dynamics of “structured” personality tests. Journal of Clinical Psychology, 1, 296–303. (no doi)

Meehl PE (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46(4), 806–834. doi: 10.1037/0022-006X.46.4.806

Meehl PE & Golden RR (1982). Taxometric methods In Kendall PC & Butcher JN (Eds.). Handbook of research methods in clinical psychology (pp. 127–181). New York: Wiley.

Messick S (1995). Standards of validity and the validity of standards in performance assessment. Educational Measurement: Issues and Practice, 14(4), 5–8.doi: 10.1111/j.1745-3992.1995.tb00881.x

Nunnally JC (1978). Psychometric theory (2nd. ed.). New York: McGraw-Hill.

Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, & Cella D (2011). Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS®): Depression, anxiety, and anger. Assessment, 18(3), 263–283. 10.1177/1073191111411667 [PubMed: 21697139]

Putnam SP, Rothbart MK, & Gartstein MA (2008). Homotypic and heterotypic continuity of fine-grained temperament during infancy, toddlerhood, and early childhood. Infant and Child Development, 17(4), 387–405. doi: 10.1002/icd.582

Reise SP, Ainsworth AT, & Haviland MG (2005). Item response theory: Fundamentals, applications, and promise in psychological research. Current Directions in Psychological Science, 14(2), 95–101. doi: 10.1111/j.0963-7214.2005.00342.x

Reise SP, & Waller NG (2009). Item response theory and clinical measurement. Annual Review of Clinical Psychology, 5, 27–48. doi: 10.1146/annurev.clinpsy.032408.153553

Rudick MM, Yam WH, & Simms LJ (2013). Comparing countdown- and IRT-based approaches to computerized adaptive personality testing. Psychological Assessment, 25(3), 769–779. doi: 10.1037/a0032541 [PubMed: 23647045]

Russell DW (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in personality and social psychology bulletin. Personality and Social Psychology Bulletin, 28(12), 1629–1646. doi: 10.1177/014616702237645

SAS Institute, Inc. (2013). SAS/ STAT software: Version 9.4. Cary, NC: SAS Institute.

Schmidt FL, Le H, & Ilies R (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. Psychological Methods, 8(2), 206–224. 10.1037/1082-989X.8.2.206 [PubMed: 12924815]

Schwartz SJ, Benet-Martínez V, Knight GP, Unger JB, Zamboanga BL, Des Rosiers SE, . . . Szapocznik J (2014). Effects of language of assessment on the measurement of acculturation: Measurement equivalence and cultural frame switching. Psychological Assessment, 26(1), 100–114. doi: 10.1037/a0034717 [PubMed: 24188146]

Simms LJ & Watson D (2007). The construct validation approach to personality scale construction In Robins RW, Fraley RC, & Krueger RF (Eds.), Handbook of research methods in personality psychology (pp. 240–258). New York: Guilford Press.

Simms LJ, Zelazny K, Williams TF, & Bernstein L (2019). Does the number of response options matter? psychometric perspectives using personality questionnaire data Psychological Assessment, 31(4), 557–566. 10.1037/pas0000648 [PubMed: 30869956]

Smith GT, & McCarthy DM (1995). Methodological considerations in the refinement of clinical assessment instruments. Psychological Assessment, 7, 300–308. doi: 10.1037/1040-3590.7.3.300

Smith GT, McCarthy DM, & Anderson KG (2001). On the sins of short-form development. Psychological Assessment, 12, 102–111. doi: 10.1037/1040-3590.12.1.102

Soto CJ, & John OP (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. Journal of Research in Personality, 68, 69–81.doi: 10.1016/j.jrp.2017.02.004

Spitzer RL, Forman JB, & Nee J (1979). DSM-III field trials: I. initial interrater diagnostic reliability. The American Journal of Psychiatry, 136(6), 815–817.doi: 10.1176/ajp.136.6.815 [PubMed: 443467]

Streiner DL (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. Journal of Personality Assessment, 80(1), 99–103.doi: 10.1207/S15327752JPA8001_18 [PubMed: 12584072]

Tackett JL, Lahey BB, van Hulle C, Waldman I, Krueger RF, & Rathouz PJ (2013). Common genetic influences on negative emotionality and a general psychopathology factor in childhood and adolescence. Journal of Abnormal Psychology, 122(4), 1142–1153. doi: 10.1037/a0034151 [PubMed: 24364617]

Tellegen A, & Waller NG (2008). Exploring personality through test construction: Development of the multidimensional personality questionnaire In Boyle GJ, Matthews G & Saklofske DH (Eds.), The SAGE handbook of personality theory and assessment, vol. 2: Personality measurement and testing (pp. 261–292) Sage Publications, Inc., Thousand Oaks, CA. doi: 10.4135/9781849200479.n1

Watson D, Clark LA, & Chmielewski M (2008). Structures of personality and their relevance to psychopathology: II. Further articulation of a comprehensive unified trait structure. Journal of Personality, 76(6), 1485–1522. 10.1111/j.1467-6494.2008.00531.x [PubMed: 19012656]

Watson D (2012). Objective tests as instruments of psychological theory and research In Cooper H (Ed.), Handbook of Research Methods in Psychology. Volume 1: Foundations, planning, measures, and psychometrics (pp. 349–369). Washington, DC: American Psychological Association.

Watson D, & Clark LA (1992). On traits and temperament: General and specific factors of emotional experience and their relation to the five-factor model. Journal of Personality, 60(2), 441–476. doi: 10.1111/j.1467-6494.1992.tb00980.x [PubMed: 1635050]

Watson D, Clark LA, Chmielewski M, & Kotov R (2013). The value of suppressor effects in explicating the construct validity of symptom measures. Psychological Assessment, 25(3), 929–941. doi: 10.1037/a0032781 [PubMed: 23795886]

Watson D, Clark LA, & Harkness AR (1994). Structures of personality and their relevance to psychopathology. Journal of Abnormal Psychology, 103, 18–31. doi: 10.1037/0021-843X.103.1.18 [PubMed: 8040477]

Watson D, Clark LA, & Tellegen A (1984). Cross-cultural convergence in the structure of mood: A Japanese replication and a comparison with U. S. findings. Journal of Personality and Social Psychology, 47, 127–144. doi: 10.1037/0022-3514.47.1.127

Watson D, Nus E, & Wu KD (2019). Development and validation of the faceted inventory of the five-factor model (FI-FFM). Assessment, 26(1), 17–44. 10.1177/1073191117711022 [PubMed: 28583005]

Watson D, O'Hara MW, Simms LJ, Kotov R, Chmielewski M, McDade-Montez E, . . . Stuart S (2007). Development and validation of the inventory of depression and anxiety symptoms (IDAS). Psychological Assessment, 19(3), 253–268. doi: 10.1037/1040-3590.19.3.253 [PubMed: 17845118]

Watson D, Stanton K, & Clark LA (2017). Self-report indicators of negative valence constructs within the research domain criteria (RDoC): A critical review. Journal of Affective Disorders, 216, 58–69. doi: 10.1016/j.jad.2016.09.065 [PubMed: 27823854]

Watson D, Stasik SM, Ellickson-Larew S, & Stanton K (2015). Extraversion and psychopathology: A facet-level analysis. Journal of Abnormal Psychology, 124(2), 432–446. doi: 10.1037/abn0000051 [PubMed: 25751628]

Watson D, Suls J, & Haig J (2002). Global self-esteem in relation to structural models of personality and affectivity. Journal of Personality and Social Psychology, 83(1), 185–197. doi: 10.1037/0022-3514.83.1.185 [PubMed: 12088125]

Zimmermann J, & Wright AGC (2017). Beyond description in interpersonal construct validation: Methodological advances in the circumplex structural summary approach. Assessment, 24(1), 3–23. doi: 10.1177/1073191115621795 [PubMed: 26685192]

Zimmerman M (1994). Diagnosing personality disorders: A review of issues and research models. Archives of General Psychiatry, 51(3), 225–245.doi: 10.1001/archpsyc.1994.03950030061006 [PubMed: 8122959]

Public Significance Statement:

Over the past 50 years, our understanding has greatly increased regarding how various psychological problems are interrelated and how they relate to various aspects of personality. In this context, this article describes a “best practice” process and relevant specific issues for developing measures to assess personality and psychological problems.

Table 1

Correlations among Selected IDAS Items (Overall Standardized Sample)

Paraphrased Item	**1**	**2**	**3**
Model 1 – Sleep problems, AIC = .46			
1. Slept less than usual	—		
2. Had trouble falling asleep	.45	—	
3. Woke up earlier than usual	.36	.34	—
4. Slept very poorly	.54	.61	.45
Model 2 – Depression, AIC = .31			
1. Felt depressed	—		
2. Did not have much of an appetite	.34	—	
3. Woke up earlier than usual	.26	22	—
4. Took a lot of effort to get going	.50	.25	.24
Model 3 – Internalizing, AIC = .29			
1. Felt dizzy or lightheaded	—		
2. Little things made me mad	.32	—	
3. Woke up earlier than usual	.24	.24	—
4. Was difficult to make eye contact	.31	.36	.24

Note. N = 8,305. IDAS-II = Expanded version of the Inventory of Depression and Anxiety Symptoms. AIC = Average interitem correlation. Target item is **bolded.**