The Importance of Being Valid

Reliability and the Process of Construct Validation

Oliver P. John
Christopher J. Soto

The truth is rarely pure and never simple. Modern life would be very tedious if it were either.

—Oscar Wilde, The Importance of Being Earnest

The empirical sciences all design measurement procedures to obtain accurate information about objects, individuals, or groups. When astronomers measure the behavior of planets and stars, when molecular biologists assay the expression of hormones in cells, when medical doctors assess how much amniotic fluid is left late in a pregnancy, they all worry whether their measurement procedures provide the kind of information they are looking for—that is, is the information generally applicable and does it capture the phenomenon they are interested in? Across all these disciplines, reliability and validity are fundamental concepts that help researchers evaluate how well their measures work.

Consider the amniotic fluid test doctors use to determine whether a fetus may be at risk late in a pregnancy. Exact measurement, as in some of the physical sciences, is not possible here; the doctor can hardly pump out all the amniotic fluid that is left in the mother’s amniotic sack to measure its volume exactly (which would indeed put the fetus at risk). Instead, medical researchers have developed an ingenious multistep procedure, using two-dimensional ultrasound images to estimate the three-dimensional volume of fluid present in the uterus. The test is administered by a technician, who produces a single number, which doctors then rely on to make critical treatment decisions. However, as doctors well know (but most patients do not), these numbers are hardly perfect. Two different technicians assessing the same woman do not always come up with the same number; the same technician assessing the same group of women on 2 subsequent days may come up with different results, and in some cases a low amniotic fluid score is obtained even though other measures confirm that the fetus is not at risk. In measurement terms, doctors worry about the degree to which their tests show interjudge agreement, retest reliability, and predictive validity, respectively.
and expectant parents (and patients more generally) should be equally worried about these issues. As this example shows, having a keen understanding of issues related to reliability and validity may be important to your own health or that of your loved ones, and this chapter is intended to provide both historical and contemporary ideas about these key measurement issues. In addition, this chapter also provides answers to questions commonly asked by students of personality psychology, such as whether an alpha reliability of .50 is high enough, why we should care whether a scale is unidimensional, and what one should do to show that a measure is valid.

Some Basic Considerations in Evaluating Measurement Procedures

These questions all illustrate the fundamental concern of empirical science with generalizability, that is, the degree to which we can make inferences from our measurements or observations in regard to other samples, items, measures, methods, outcomes, and so on (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; see also Brennan, 2001). If we cannot make such generalizations, our measurements are obviously much less useful than if we can provide explicit evidence for generalizability.

Good measurement implies not only that we can reproduce or replicate a certain score, but that we can trust that the measurement has a particular meaning—we want to be able to make inferences about other variables that interest us. In the amniotic fluid test example, the volumetric measurements would be useless if they failed to help doctors predict which babies are at risk and should be delivered soon. Another basic idea is that all measures—self-reports, observer ratings, even physiological measures—are prone to errors and that we cannot simply assume that a single measurement will generalize. Any one measurement may be distorted by numerous sources of error (e.g., the medical technician may have made a human error, or the position of the fetus and the umbilical cord may have been unusual, etc.), and the resulting observation (or score) is therefore related only imperfectly to what we want to measure, namely, the risk to the baby. To counteract this limitation of single measurements, psychologists aim to obtain multiple measurements (e.g., across different stimuli, experimenters, or observers) and then aggregate them into a more generalizable composite score.

Personality Measurement: Formulating and Evaluating Models in the Psychometric Tradition

What is measurement and how may it be defined? Early on, Stevens (1951) suggested that measurement is the assignment of numbers to objects or events according to rules. More recently, Dawes and Smith (1985) and others have argued that measurement is best understood as the process of building models that represent phenomena of interest, typically in quantitative form. The raw data in personality research initially exist only in the form of minute events that constitute the ongoing behavior and experience of individuals. Judd and McClelland (1998, pp. 3–4) suggest that measurement is the process by which these infinitely varied observations are reduced to compact descriptions or models that are presumed to represent meaningful regularities in the entities that are observed . . . Accordingly, measurement consists of rules that assign scale or variable values to entities to represent the constructs that are thought to be theoretically meaningful. (emphasis added)

Like most models, measurement models (e.g., tests or scales) have to be reductions or simplifications to be useful. Although they should represent the best possible approximation of the phenomena of interest, we must expect them, like all “working models,” to be eventually proven wrong and to be superseded by better models. For this reason, measurement models must be specified explicitly so that they can be evaluated, disconfirmed, and improved. Moreover, we should not ask whether a particular model is true or correct; instead, we should build several plausible alternative models and ask, Given everything we know, which models can we rule out and which model is currently the best at representing our data? Or, even more clearly, which model is the least wrong? This kind of comparative model testing (e.g., Judd, McClelland, & Culhane, 1995) is the best strategy for evaluating and improving our measurement procedures.
Organization of This Chapter

This chapter is organized into three major parts. We begin with historically early conceptions of reliability, then move on to increasingly complex views that emphasize the construct validation process, and finally consider model testing as an integrative approach. Specifically, we first consider issues traditionally discussed under the heading of reliability, review still-persistent "types" of reliability coefficients, then suggest generalizability theory as a broader perspective, and finally discuss in some detail the problems and misuses of coefficient alpha, the most commonly used psychometric index in personality psychology. Second, we discuss five kinds of evidence that are commonly sought in the process of construct validation, which we view as the most crucial issue in psychological measurement. In the third part, we consider model testing in the context of construct validation; following a brief introduction to measurement models in structural equation modeling (SEM), we discuss an empirical example that presents the issue of dimensionality as an aspect of structural validity.

From a Focus on "Reliability Coefficients" to Generalizability Theory

As our introductory examples illustrate, most measurement procedures in psychology and other empirical disciplines are subject to "error." In personality psychology, the observations, ratings, or judgments that constitute the measurement procedure are typically made by humans who are subject to a wide range of frailties. Research participants may become careless or inattentive, bored or fatigued, and may not always be motivated to do their best. The particular conditions and point in time when ratings are made or recorded may also contribute error. Further errors may be introduced by the rating or recording forms given to the raters; the instructions, definitions, and questions on these forms may be difficult to understand or require complex discriminations, again entering error into the measurement.

These various characteristics of the participants, the testing situation, the test or instrument, and the experimenter can all introduce measurement error and thus affect what has traditionally been called reliability. Reliability refers to the consistency of a measurement procedure, and indices of reliability all describe the extent to which the scores produced by the measurement procedure are reproducible.

Reliability in Classical Test Theory

Issues of reliability have traditionally been treated within the framework of classical test theory (Gulliksen, 1950; Lord & Novick, 1968). If a given measurement $X$ is subject to error $e$, then the measurement without the error, $X - e$, would represent the accurate or "true" measurement $T$. This seemingly simple formulation, that each observed measurement $X$ can be partitioned into a true score $T$ and measurement error $e$, is the fundamental assumption of classical test theory. Conceptually, each true score represents the mean of a very large number of measurements of a specific individual, whereas measurement error represents all of the momentary variations in the circumstances of measurement that are unrelated to the measurement procedure itself. Such errors are assumed to be random (a rather strong assumption, to which we return later), and it is this assumption that permits the definition of error in statistical terms.

Conceptions of reliability all involve the notion of repeated measurements, such as over time or across multiple items, observers, or raters. Classical test theory has relied heavily on the notion of parallel tests—that is, two tests that have the same mean, variance, and distributional characteristics and correlate equally with external variables (Lord & Novick, 1968). Under these assumptions, true score and measurement error can be treated as independent. It follows that the variance of the observed scores equals the sum of the variance of the true scores and the variance of the measurement error:

$$\sigma^2_X = \sigma^2_T + \sigma^2_e$$

Reliability can then be defined as the ratio of the true-score variance to the observed-score variance, which is equivalent to 1 minus the ratio of error variance to observed-score variance:

$$r_{xx} = \frac{\sigma^2_T}{\sigma^2_X} = 1 - \frac{\sigma^2_e}{\sigma^2_X}$$

If there is no error, then the ratio of true-score
variance to total variance (and hence reliability) would be 1; if there is only error and no true-score variance, then this ratio (and hence reliability) would be 0.

### Costs of Low Reliability, and Correcting Observed Correlations for Attenuation Due to Low Reliability

Classical test theory (Lord & Novick, 1968) suggests that researchers ought to work hard to attain high reliabilities because the reliability of a measure constrains how strongly that measure may correlate with another variable (e.g., an external criterion). If error is truly random, as classical test theory assumes, the upper limit of the correlation for a measure is not 1.0 but the square root of its reliability (i.e., the correlation of the measure with itself). Thus, the true correlation between the measure and another variable may be underestimated (i.e., attenuated) when reliability is inadequate. In other words, low reliability comes at a cost.

Students sometimes ask questions like "My scale has a reliability of .70—isn't that good enough?" and are frustrated when the answer is, "That depends." Although it would be quite convenient to have a simple cookbook for measurement decisions, there is no minimum or optimum reliability that is necessary, adequate, or even desirable in all contexts. Over the years a convention seems to have evolved, often credited to Nunnally (1978), that regards "reliabilities of .7 or higher" (p. 245) as sufficient. However, a reliability of .70 is not a benchmark every measure must pass. In the words of Pedhazur and Schmelkin (1991),

"Does a .5 reliability coefficient stink? To answer this question, no authoritative source will do. Rather, it is for the user to determine what amount of error variance he or she is willing to tolerate, given the specific circumstances of the study." (p. 110, emphasis in original)

We wondered, then, is there something useful in the widely shared view of .70 as the "sweet spot" of reliability? To find out, we examined the relative costliness of various levels of reliability, as presented in Table 27.1. We derived the numbers in Table 27.1 by rewriting the formula traditionally used to correct observed correlations for attenuation due to unreliability (Cohen, Cohen, West, & Aiken, 2003; Lord & Novick, 1968) and solving it for the observed correlations. This equation estimates the expected observed correlation (r_{XY}), given the true correlation between the constructs measured by X and Y (p_{XY}) and the geometric average of the two measures' reliabilities (r_{XX} and r_{YY})

\[
r_{XY} = \frac{p_{XY} \sqrt{r_{XX} \cdot r_{YY}}}{\sqrt{r_{XX} + r_{YY} - 2 \cdot r_{XX} \cdot r_{YY}}}
\]

As shown in Table 27.1, if X and Y both have a high reliability of .90 (or an average reliability of .90), then the losses are modest. For example, a true correlation of .70 (an unusually large effect size) would result in an observed correlation of .63 and a true correlation of .30 (a common effect size) would still result in an observed correlation of .27. In short, the loss due to unreliability would be quite small, with observed correlations being 90% of the true correlations—that is, only 10% lower. At an average reliability of .70 (a common situation in personality research), the losses would be more pronounced: A true correlation of .70 would be reduced to .49, and a true correlation of .30 to .14, with observed correlations being only 70% of the true correlations—a loss of 30%. At an average reliability of .50, the losses would be drastic: A true correlation of .70 would become a mere .35, and a true correlation of .30 would be reduced to .15—a 50% loss.

It is easy to see that with small sample sizes and true effect sizes typically in the .15–.40 range, discovering real effects with unreliable measurements becomes increasingly difficult.
Making the costs of reliabilities in the .50 range prohibitive. For example, with a sample of 100 participants and true correlations in the .30 range, an average reliability of .70 is barely large enough to observe statistically significant correlations. If we assume that this scenario is quite common in the field, then the benchmark reliability of "70 or above" makes some sense; certainly one would not want to accept reliabilities lower than .70 if that means being unable to detect expected correlations in the .30 range.

However, the costs of reliabilities lower than .70 can be at least partially offset. As Table 27.1 shows, if true-correlation sizes are large (or sample sizes are large), lower reliabilities are more easily tolerated because expected effects would still be detected at conventional significance levels. Nonetheless, it must be emphasized that even under these favorable conditions, the true effect sizes will be severely underestimated—a grave disadvantage, given that obtaining replicable estimates of the size of a correlation is now deemed much more important than its statistical significance in any one sample (see Fraley & Marks, Chapter 9, this volume).

Researchers sometimes use reliability indices to correct observed correlations between two measures for attenuation due to unreliability. The correction formula (Cohen et al., 2003; Lord & Novick, 1968) involves dividing the observed correlation by the square root of the product of the two reliabilities:

$$\rho_{XY} = \frac{r_{XY}}{\sqrt{r_{XX}r_{YY}}}$$

This correction expresses the size of the association relative to the maximum correlation attainable, given the imperfect reliabilities of the two measures. This kind of correction is sometimes used to estimate the true correlation between the latent constructs underlying the measures (see also the section on SEM below), thus indicating what the observed correlation would be if both constructs were assessed with perfect reliability. Correction for attenuation can also be useful when researchers want to compare effect sizes across variables or studies that use measures of varying reliabilities, as in meta-analyses (see Roberts, Kuncel, Viechtbauer, & Bogg, Chapter 36, this volume). Another application is in contexts where researchers want to distinguish the long-term stability of personality and attitudes from the reliability of measurement.

However, the ease with which this correction is made should not be seen as a license for sloppy measurement. In many situations, low reliability will create problems for estimating effect sizes, testing hypotheses, and estimating the parameters in structural models—problems that cannot be overcome by simply correcting for attenuation due to unreliability. This is especially true in multivariate applications, such as multitrait-multimethod matrices (discussed below), where unequal reliabilities might bias conclusions about convergent and discriminant validity (West & Finch, 1997). In general, then, researchers are well-advised to invest the time and effort needed to construct reliable measures and consult Table 27.1 to gauge the amount of measurement error that they are willing to tolerate, given the goals of their research.

Evidence for Reliability: Traditional Types of "Reliability Coefficients"

The three most common procedures to assess reliability are shown in Table 27.2: internal consistency (or split-half), retest (or stability), and interrater agreement designs. The American Psychological Association (APA) committee on psychological tests articulated these types of designs to clarify that "reliability is a generic term referring to many types of evidence" (American Psychological Association, 1954, p. 28). Clearly, the different study designs in Table 27.2 assess rather different sources of error. Internal consistency procedures offer an estimate of error associated with the particular selection of items; error is high (and internal consistency is low) when items are heterogeneous in content and lack content saturation and when respondents change how they respond to items designed to measure the same characteristic (e.g., owing to fatigue). Retest (or stability) designs estimate how much responses vary within individuals across time and situation, thus reflecting error due to differences in the situation and conditions of test administration or observation.1 Interrater or interjudge agreement designs estimate how much scores vary across judges or observers (see von Eye & Mun, 2003), thus reflecting error due to disagreements between raters and to individual differences among raters in response styles, such as the way they scale their responses. It is important to note that, as Table 27.2 shows, there are several different reliabil-
Moving beyond the Classical Conception of Reliability: Generalizability Theory

The distinctions among “types of reliability coefficients” had a number of unfortunate consequences. First, what had been intended as heuristic distinctions became reified as the stability coefficient or the alpha coefficient even though the notion of reliability was intended as a general concept. Second, the classification itself was too simple, equating particular kinds of reliability evidence with only one source of error and resulting in a restrictive terminology that cannot fully capture the broad range and combination of multiple error sources that are of interest in most research and measurement applications (e.g., Shavelson, Webb, & Rowley, 1989). For example, as we show in Table 27.2, retest reliability involves potential changes over time in both the research participants and the testing conditions (e.g., prior to vs. just after September 11, 2001).

Third, the types-of-reliability approach masked a major shortcoming of classical test theory: If all these measures were indeed parallel and all errors truly random, then all these approaches to reliability should yield the same answer. Unfortunately, they do not, because reliability depends on the particular facet of generalization being examined (Cronbach, Rajaratnam, & Gleser, 1963). For example, to address the need for superbrief scales of the Big Five trait domains for use in surveys and experimental contexts, researchers have recently constructed scales consisting of only two or four items each ( Gosling, Rentfrow, & Swann, 2003; Rammstedt & John, 2006, 2007). These items were not chosen to be redundant, but instead to represent the broad Five domains, as well as to balance scoring by including both true-scored and false-scored items. Not surprisingly, the resulting scales had very low internal consistency (alpha) reliabilities. Does this mean that these scales are generally unreliable? Not at all. They do show impressive reliability when other facets of generalizability are considered. For example, with less than one-quarter of the items of the

Table 27.2. Traditional Reliability Coefficients: Study Design, Statistics, Sources of Error, and Facets of Generalizability

<table>
<thead>
<tr>
<th>Traditional reliability coefficient</th>
<th>Study design</th>
<th>Reliability statistic</th>
<th>Major sources of error</th>
<th>Facet of generalizability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal consistency</td>
<td>Measure participants at a single time across multiple items</td>
<td>Cronbach's coefficient alpha (split-half correlations rarely used today)</td>
<td>Heterogeneous item content; participant fatigue</td>
<td>Items</td>
</tr>
<tr>
<td>Retest</td>
<td>Measure the same participants across two or more occasions or times using the same set of items</td>
<td>Correlation between participants' scores at the two times</td>
<td>Change in participants' responses; change in measurement situation</td>
<td>Occasions</td>
</tr>
<tr>
<td>Interrater (interjudge)</td>
<td>Obtain ratings of a set of stimuli (individuals, video recordings, transcribed interviews) from multiple raters (e.g., observers, coders) at one time</td>
<td>1. Mean pairwise interrater agreement correlation (for reliability of a typical single rater) 2. Cronbach's coefficient alpha (for reliability of the mean rating) 3. Cohen's kappa (for agreement of categorical ratings)</td>
<td>Disagreement among raters; variation in raters' response styles</td>
<td>Raters</td>
</tr>
</tbody>
</table>

Internal consistency indices, and not all of them are based on correlations; therefore, different criteria for evaluating the reliability of a measure will be needed. Values approaching 1.0 are not expected for all reliability indices—for example, Cohen's kappa, which measures agreement among categorical judgments (see, e.g., von Eye & Mun, 2005).
Facet of generalizability

Items

Occasions

Raters

full-length 44-item Big Five Inventory (BFI; see John & Srivastava, 1999), the BFI-10 scales can still represent the content of the full scales with an average part-whole correlation of .83; 6-week retest correlations average .75. In short, different types of reliability have conceptually distinct meanings that do not necessarily cohere.

Therefore, the American Psychological Association (e.g., 1985) recommended in subsequent editions of the Standards for Educational and Psychological Testing that these distinctions and terminology be abolished and replaced by the broader view advocated by generalizability theory (Cronbach et al., 1963). Regrettably, however, practice has not changed much over the years, and generalizability theory has not fully replaced these more simplistic notions. Note that the last column in Table 27.2 spells out the facet of generalizability that is being varied and studied in each of these generalizability designs.

Generalizability theory holds that we are interested in the "reliability" of an observation or measurement because we wish to generalize from this observation to some other class of observations. For example, Table 27.2 shows that a concern with interjudge reliability may actually be a concern with the question of how accurately we can generalize from one judge to another (pairwise agreement) or from a given set of judges to another set (generalizability of aggregated or total scores). Or we may want to know how well scores on an attitude scale constructed according to one set of procedures generalize to another scale constructed according to different procedures. Or we may want to test how generalizable is a scale originally developed in English to a Chinese language and cultural context.

All these facets of generalizability represent legitimate research concerns (see the later section on construct validation), and they can be studied systematically in generalizability designs, both individually and together. These designs allow the researcher to deliberately vary the facets that potentially influence observed scores and estimate the variance attributable to each facet (Cronbach et al., 1972). In other words, whereas classical test theory tries to estimate the portion of variance that is attributable to "error," generalizability theory aims to estimate the extent to which specific sources of variance contribute to test scores under carefully defined conditions. Thus, instead of the traditional reliability coefficients listed in Table 27.2, we should use more general estimates, such as intraclass correlation coefficients (Shrout & Fleiss, 1979), to probe particular aspects of the dependability of measures. For example, intraclass coefficients can be used to index the generalizability of one set of judges to a universe of similar judges (von Eye & Mun, 2005).

Generalizability theory should hold considerable appeal for psychologists because the extent to which we can generalize across items, instruments, contexts, groups, languages, and cultures is crucial to the claims we can make about our findings. Despite excellent and readable introductions (e.g., Brennan, 2001; Shavelson et al., 1989), generalizability theory is still not used as widely as it should be. A recent exception is the flourishing research on the determinants of consensus among personality raters (e.g., Kenny, 1994; see also John & Robins, 1993; Kashy & Kenny, 2000; Kwan, John, Kenny, Bond, & Robins, 2004).

Generalizability theory is especially useful when data are collected in nested designs and multiple facets may influence reliability, as illustrated by King and Figueredo's (1997) research on chimpanzee personality differences. The investigators collected ratings of chimpanzees, differing in age and sex (subject variables) on 40 traits (stimulus variables) at several different zoos (setting variables), from animal keepers familiar with the animals to varying degrees (observer variables). They then used a generalizability design to show how these facets affected agreement among the judges. It is unfortunate that generalizability theory, as well as Kenny's (1994) social relations model, have been perceived as "technical." With clear and accessible introductions available, it is high time that these important approaches to variance decomposition achieve greater popularity with a broader group of researchers.

Coefficient Alpha: Personality Psychology's Misunderstood Giant

Cronbach's (1951) coefficient alpha is an index of internal consistency that has become the default reliability index in personality research. Any recent issue of a personality journal will show that alpha is the index of choice when researchers want to claim that their measure is reliable. Often it is the only reliability evidence considered, contrary to the recommendations
in the Standards (American Psychological Association, 1983).

We suspect that alpha has become so ubiquitous because it is easy to obtain and compute. Alpha does not require collecting data at two different times from the same subjects, as retest reliability does, or the construction of two alternate forms of a measure, as parallel-form reliability (now rarely used) would require. Alpha is a "least effort" reliability index—it can be used as long as the same subjects responded to multiple items thought to indicate the same construct. And, computationally, SPSS and other statistical packages now allow the user to view the alpha of many alternative scales formed from any motley collection of items with just a few mouse clicks. Unfortunately, whereas alpha has many important uses, it also has important limitations—long known to methodologists, these limitations are less well appreciated by researchers and thus worth reviewing in some detail.

**Alpha Is Determined by Both Item Homogeneity (Content Saturation) and Scale Length**

The alpha coefficient originated as a generalization of split-half reliability, representing the corrected mean of the reliabilities computed from all possible split-halves of a test. As such, alpha is a function of two parameters: (1) the homogeneity or interrelatedness of the items in a test or scale (as indexed by the mean intercorrelation of all the items on the test, \( r_{ij} \)) and (2) the length of the test (as indexed by the number of items on the test, \( k \)). The formula is

\[
\alpha = \frac{kr_0}{k r_0 + (1 - r_0)}
\]

Conceptually, note that the term on the top of the fraction allows alpha to increase as the number of items on the scale goes up and as the mean intercorrelation between the items on the scale increases. However, to constrain alpha to a range from 0 to 1, the same term repeats at the bottom of the fraction plus a norming term (1 - mean \( r_{ij} \)) that increases the divisor as the mean interitem correlation decreases. If that interitem correlation were indeed 1, the norming term would reduce to 0 and alpha would be its maximum of 1.0 regardless how many items were on the scale. Conversely, if the interitem correlation were 0, the numerator would become 0 and so would alpha, again regardless of the number of items on the scale. Finally, alpha is defined meaningfully only for \( K \geq 2 \) because at least 2 items are needed to compute the mean interitem correlation required for the formula. Consider a questionnaire scale with 9 items and mean \( r_{ij} = .42 \): alpha would be \((9 \times .42)/[9 \times .42 + (1 - .42)] = 3.78/(3.78 + .58) = .87\).

What exactly does alpha mean, then? An alpha of .87 means, in plain English, that the total score derived from aggregating these 9 items would correlate .87 with the total score derived from aggregating another (imaginary) set of 9 equivalent items; that is, alpha captures the generalizability of the total score from one item set to another item set. The term internal consistency is therefore a misleading label for alpha: The homogeneity or interrelatedness of the items on the scale and the length of the scale have been aggregated and thus integrated into the total score, and the generalizability of this total score (i.e., alpha) can therefore no longer tell us anything concrete about the internal structure or consistency of the scale. The hypothetical data presented in Table 27.3 were designed to make these points as concrete and vivid as possible.

Table 27.3 shows the interitem correlation matrices for three hypothetical questionnaire scales. Following Schmitt (1996), we constructed our examples in correlational (rather than covariance) terms for ease of interpretation. Scale A is the one we just considered, with 9 items, a mean interitem correlation of .42, and an alpha of .87. Scale B has 6 items, and it has the same alpha of .87 as Scale A. But note that Scale B attained that alpha in a rather different way. Scale B has 3 fewer items, but that deficiency is offset by the greater homogeneity or content saturation of its items: They are more highly intercorrelated (mean \( r_{ij} = .52 \)) than are the 9 items of Scale A (mean \( r_{ij} = .42 \)).

This example illustrates the idea that test length can compensate for lower levels of interitem correlation, an idea that is formalized in the Spearman-Brown prophecy formula, which specifies the relation between test length and reliability (see, e.g., Lord & Novick, 1968). For any mean interitem correlation, the formula computes how many items are needed to achieve a certain level of alpha. Figure 27.1 shows this relation for mean interitem correlations of .20, .40, .60, and .80. Three points are worth noting. First, the alpha reliability of the total scale always increases as the number of
TABLE 27.3. Interitem Correlation Matrix for Three Hypothetical Scales with Equal Coefficient Alpha Reliability

| Scale A: 9 items, mean interitem correlation = .42, \( \alpha = .87 \) |
|-------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                   | 1               | 2               | 3               | 4               | 5               | 6               | 7               | 8               | 9               |
| 1                  | -               | .42             | -               | .42             | .42             | .42             | .42             | .42             | .42             |
| 2                  | .42             | -               | .42             | -               | .42             | .42             | .42             | .42             | .42             |
| 3                  | .42             | .42             | -               | .42             | .42             | .42             | .42             | .42             | .42             |
| 4                  | .42             | .42             | .42             | -               | .42             | .42             | .42             | .42             | .42             |
| 5                  | .42             | .42             | .42             | .42             | -               | .42             | .42             | .42             | .42             |
| 6                  | .42             | .42             | .42             | .42             | .42             | -               | .42             | .42             | .42             |
| 7                  | .42             | .42             | .42             | .42             | .42             | .42             | -               | .42             | .42             |
| 8                  | .42             | .42             | .42             | .42             | .42             | .42             | .42             | -               | .42             |
| 9                  | .42             | .42             | .42             | .42             | .42             | .42             | .42             | .42             | -               |

| Scale B: 6 items, mean interitem correlation = .52, \( \alpha = .87 \) |
|-------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                   | 1               | 2               | 3               | 4               | 5               | 6               |
| 1                  | -               | .52             | -               | .52             | -               | .52             |
| 2                  | .52             | -               | .52             | .52             | -               | .52             |
| 3                  | .52             | .52             | -               | .52             | .52             | -               |
| 4                  | .52             | .52             | .52             | -               | .52             | .52             |
| 5                  | .52             | .52             | .52             | .52             | -               | .52             |
| 6                  | .52             | .52             | .52             | .52             | .52             | -               |

| Scale C: 6 items, mean interitem correlation = .52, \( \alpha = .87 \) |
|-------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                   | 1               | 2               | 3               | 4               | 5               | 6               |
| 1                  | -               | .70             | -               | .70             | -               | .70             |
| 2                  | .70             | -               | .70             | -               | .70             | -               |
| 3                  | .70             | .70             | -               | .70             | .70             | -               |
| 4                  | .40             | .40             | .40             | -               | .70             | .70             |
| 5                  | .40             | .40             | .40             | .70             | -               | .70             |
| 6                  | .40             | .40             | .40             | .70             | .70             | -               |

Note: CFA analyses showed that for Scale A, all items load .648 on a single factor; fit is perfect. For Scale B, all items load .721 on a single factor; fit is perfect. In contrast, Scale C is not unidimensional; for the one-factor model, all items load .721 and standardized root mean residual (RMR) is only .124. For two-factor models for Scale C, all items load .837 on their factor; the interfactor correlation is .571 and fit is perfect.
Clearly, the responses to these 6 items are a function of not one, but two, factors: Items 1, 2, and 3 correlate much more substantially with each other (mean \( r = .7 \)) than they correlate with items 4, 5, and 6, which in turn correlate more highly among themselves (mean \( r = .4 \)). Alpha completely disguises this rather important difference between Scale C and B.

Because alpha cannot address it, unidimensionality needs to be established in other ways, as we describe in later sections of this chapter on structural validity and on model testing; there we also discuss factor analyses of the example data in Table 27.3. Here, it is important to emphasize that the issue of error (or unreliability) present in an item is separate from the issue of multidimensionality. In other words, unidimensionality does not imply lower levels of error (i.e., unreliability), and multidimensionality does not imply higher levels of error. Once we know that a test is unidimensional, can we go ahead and still use alpha as a reliability index? The answer is no. The reliability of a multidimensional scale can be estimated only through parallel forms, which must have the same factor structure (Cronbach, 1947, 1951). In fact, if the test is not unidimensional, then alpha underestimates reliability (see Schmitt, 1996, for an example). Thus, if a test is found to be mul-

**FIGURE 27.1.** Cronbach’s coefficient alpha reliability as a function of the number of items on a scale \((k)\) and the mean of the correlations among all the items \((\text{mean } r_{ij})\); see the text for the formula used to generate this graph.
Reliability and Construct Validation

471

tidimensional, one should score two unidimensional subscales and then use alpha to index their reliabilities separately.2

Evaluating the Size of Alpha

According to Classical Test Theory, increasing alpha can have only beneficial effects. As discussed previously, higher reliability means that a greater proportion of the individual differences in measurement scores reflect variance in the construct being assessed (as opposed to error variance), thus increasing the power to detect significant relations between variables. In reality, however, instead of assuming that a bigger alpha coefficient is always better, alpha must be interpreted in terms of its two main parameters—interitem correlation and scale length—and in the context of how these two parameters fit the definition of the particular construct to be measured. In any one context, a particular level of alpha may be too high, too low, or just right.

Alpha and Item Redundancy

Consider a researcher who wants to measure the broad construct of neuroticism—which includes anxiety, depression, and hostility as more specific facets (see Figure 27.2). The researcher has developed a scale with the following items: "I am afraid of spiders," "At times I think I am not anxious around creepy-crawly things," "I am not bothered by insects" (reverse scored), and "Spiders tend to make me nervous." Note that these items are essentially paraphrases of each other and represent the same item content (arachnophobia or being afraid of insects) stated in slightly different ways. Cattell (1972) considered scales of this kind to be "bloated specifics"—they have high alphas simply because the item content is extremely redundant and the resulting interitem correlations are very high. Thus, alphas in the high .80's or even .90's, especially for short scales, may not indicate an impressively reliable scale but instead signal redundancy or narrowness in item content. Such measures are susceptible to the so-called attenuation paradox: Increasing the internal consistency of a test beyond a certain point will not enhance validity and may even come at the expense of validity when the added items emphasize one narrow part of the construct over other important parts.

An example of a measure with high item redundancy is the 10-item Rosenberg (1979) self-esteem scale, which has alphas approaching .90 and some interitem correlations approaching .70 (Gray-Little, Williams, & Hancock, 1997; Robins, Hendin, & Trzesniewski, 2001). Not surprisingly, some of these items turn out to be almost synonymous, such as "I certainly feel useless at times" and "At times I think I am no good at all." Although such redundant items increase alpha, they do not add unique (and thus incremental) information and can often be omitted in the interest of efficiency, suggesting that the scale can be abbreviated without much loss of information (see, e.g., Robins et al., 2001). More recently, considerable item redundancy was noted by the authors of the original and revised Experiences in Close Relationships questionnaires (ECR and ECR-R; Brennan, Clark, & Shaver, 1998; Fraley, Waller, & Brennan, 2000). One example of a redundant item pair, from the ECR anxiety scale, is "I worry about being abandoned" and "I worry a fair amount about losing my partner." A second example, from the ECR avoidance scale, is "Just when my partner starts to get close to me, I find myself pulling away" and "I want to get close to my partner, but I keep pulling back." We have observed interitem correlations in excess of .70 for each of these pairs (Soto, Gorchoff, & John, 2006).

The fear-of-insects items on the hypothetical neuroticism scale illustrate how easy it is to boost alpha by writing redundant items. However, unless one is specifically interested in insect phobias, this strategy is not very useful. The narrow content representation (i.e., high content homogeneity) would make this scale less useful as a measure of the broader construct of neuroticism. Although the scale may predict the intensity of emotional reactions to spiders with great precision (or fidelity), it is less likely to relate to anything else of interest because of its very narrow bandwidth. Conversely, broadband measures (e.g., a neuroticism scale) can predict a wider range of outcomes or behaviors but generally do so with lower fidelity. This phenomenon is known as the bandwidth-fidelity tradeoff (Cronbach & Gleser, 1957) and has proven to be of considerable importance in many literatures, including personality traits (Epstein, 1980; John, Hampson, & Goldberg, 1991) and attitudes (Eagly & Chaiken, 1993; Fishbein & Ajzen, 1974). In general, predictive accuracy is maximized when the trait or attitude serving as the pre-
ANALYZING AND INTERPRETING PERSONALITY DATA

The close connection between the hierarchical level of the construct to be measured and the content homogeneity of the items is illustrated in Figure 27.2. Anxiety, depression, and hostility are three trait constructs that tend to be positively intercorrelated and together define the broader construct of neuroticism (e.g., Costa & McCrae, 1992). (Our initial example of the insect phobia scale might be represented as an even lower-level construct, one of many more specific components of anxiety.) Consider now the anxiety scale on the left side of Figure 27.2. Because its six items represent a narrow range of content (e.g., being fearful, nervous, and worrying), item content should be relatively homogeneous, leading to a reasonably high mean interitem correlation and, with 6 items, a reasonable alpha reliability. Similar expectations should hold for the six-item depression and hostility scales.

Researchers rarely publish or even discuss interitem correlations, and so far we have focused on hypothetical data to illustrate general issues. How high are typical interitem correlations on personality questionnaire scales? Are they closer to .70, indicating a high level of redundancy, or to .30, suggesting more modest overlap? Table 27.4 provides real data from a sample of University of California–Berkeley undergraduates (N = 649); for this illustration, we used their responses to a subset of 12 neuroticism items selected from Costa and McCrae’s (1992) NEO Personality Inventory—Revised (NEO-PI-R) anxiety (A) and depression (D) facet scales. In the NEO-PI-R, each of the Big Five personality domains is defined by 6 “facet” scales that each have 8 items; the resulting 48-item Big Five scales are very long and thus all have alphas exceeding .90.

We examine the reliability of the anxiety and depression facet scales first. Consider the relevant within-facet interitem correlations that are set in bold in Table 27.4. All these correlations were positive and significant; their mean was .38 for the 6 anxiety items and .39 for the 6 depression items, as shown in Figure 27.2. That is, even for these lower-level facet scales, the items on the scale correlated only modestly with each other. With these moderate interitem correlations, the 6-item facet scales attained alphas of .78 for anxiety and .79 for depression (Figure 27.2).

![Figure 27.2: Illustration of hierarchical relations among constructs and homogeneity of item content. A general neuroticism factor and more specific 6-item facets of anxiety (A), depression (D), and hostility (H), the results of internal consistency (alpha) analyses for each scale are shown in parentheses (N = 649).](image-url)
Now consider the between-facet interitem correlations, set in regular type; their mean was .28, lower on average than the (bold) within-facet correlations. This is as expected: Although all 12 of these neuroticism items should intercorrelate, the cross-facet (or discriminant) correlations of anxiety items with depression items should be lower than the within-facet (or convergent) correlations. This convergent-discriminant pattern generally held for the individual items, but there was at least one problematic item, A4 ("I often feel tense and jumpy"). Consider the italicized interitem correlations for A4. Three of its within-facet correlations (with A1, A2, and A3) were lower than four of its cross-facet correlations (with D1, D2, D3, and D6), indicating that this item did not clearly differentiate anxiety from depression.

The hierarchical-structure model in Figure 27.2 implies that the facet scales should cohere as components of the broader neuroticism domain but also differentiate anxious from depressed mood. How high, then, should facet intercorrelations be? In our student sample, the anxiety scale and depression scale correlated .78; even when corrected for attenuation due to unreliability, the estimated true correlation of .74 remained clearly below 1.0. More generally, using the full-length 8-item NEO-PI-R facet scales, the mean correlation between facets within the same superordinate Big Five domain was .40, indicating that same-domain facet scales share some common variance but also retain some uniqueness. This moderate correlation also reminds us that superordinate dimensions can be measured reliably with relatively heterogeneous “building blocks” as long as there are enough such blocks—six facets per Big Five domain in the case of the NEO-PI-R.

Finally, scales for broadband constructs like neuroticism must address issues of item heterogeneity. Consider a 6-item neuroticism scale (shown on the right side of Figure 27.2), consisting of two anxiety items, A1 and A2; two depression items, D1 and D2; and two hostility items, H1 and H2. As compared with the lower-level facet scales, the item content on this superordinate scale is much more heterogeneous, which should lead to a lower mean interitem correlation and thus a lower alpha, given that scale length is constant at 6 items. Indeed, the analyses in our student sample bear out this prediction; the mean interitem correlation was only .28 and alpha .70.

One implication of Figure 27.2 is that if one wants to measure broader constructs such as neuroticism, one should probably include a larger number of items to compensate for the greater content heterogeneity. For example,
one might use all 18 items, from A1 to H6, to measure the superordinate neuroticism construct defined on the left side of Figure 27.2. As one would expect, the mean interitem correlation for the 18-item scale was .26, just as low as that for the 6-item neuroticism scale, but this longer scale had an impressive alpha of .87. As we discuss next, however, the strategy of increasing alpha by increasing scale length can be taken too far.

Alpha and Scale Length
Whereas scales with unduly redundant item content have conceptual limitations, scales that bolster alpha by including a great many items have practical disadvantages. Overlong scales or assessment batteries can produce respondent fatigue and therefore less reliable responses (e.g., Burisch, 1984). Lengthy scales also consume an inordinate amount of participants’ time, making it likely that researchers will use them only if their interests lie solely with the construct being measured. Recognition of these disadvantages has led to a growing number of very brief measures.

Indeed, for very brief scales, alpha may not be a sensible facet of generalizability at all. For example, in our discussion of generalizability theory, we noted that very brief Big Five scales did not have high interitem correlations, because the items were chosen to represent very broad constructs as comprehensively as possible (e.g., Rammstedt & John, 2006, in press). Not surprisingly, these scales had paltry alphas; more important, they showed substantial retest reliability and predicted the longer scales that they were designed to represent quite well, findings that also hold for other innovative short measures, such as the single-item self-esteem scale (Robins et al., 2001).

Resisting the Temptation of Too-High Alphas
What can be done to prevent the construction of scales whose internal consistencies are too high? Some rules of thumb can serve as a start. We suggest that scale developers review all pairs of items that intercorrelate .60 or higher to decide whether one item in the pair may be eliminated. Ultimately, however, there is no foolproof empirical solution; as scale developers, only good judgment can save us from the siren song of inappropriately maximized alphas. Specifically, we must keep in mind that, beyond a certain point, the length of a scale will be inversely related to its usefulness for many researchers—researchers are operating under constraints on the recruitment and assessment of participants. We must also control our desire to maximize interitem correlations by way of item redundancy. We can achieve this by making sure that, throughout the scale development process, the breadth of the construct we intend to measure is reflected in the breadth of the scale items we intend to measure it with.

Item Response Theory
Classical test theory has also been criticized by advocates of item response theory (IRT; e.g., Embretson, 1996; Embretson & Reise, 2000; Mellenbergh, 1996). In classical theory, the characteristics of the individual test taker and those of the test cannot be separated (Hambleton, Swaminathan, & Rogers, 1991). That is, the person’s standing on the underlying construct is defined only in terms of responses on the particular test; thus, the same person may appear quite liberal on a test that includes many items measuring extremely conservative beliefs but quite conservative on a test that includes many items measuring radical liberal beliefs. The psychometric characteristics of the test also depend on the particular sample of respondents; for example, whether a belief item from a conservatism scale reliably discriminates high and low scorers depends on the level of conservatism of the sample, so that the same test may work well in a very liberal student sample but fail to make reliable distinctions among the relatively more conservative respondents in an older sample. In short, classical test theory does not apply if we want to compare individuals who have taken different tests measuring the same construct or if we want to compare items answered by different groups of individuals.

Another limitation of classical test theory is the assumption that the degree of measurement error is the same for all individuals in the sample—an implausible assumption, given that tests and items differ in their ability to discriminate among respondents at different levels of the underlying construct (Lord, 1984). Moreover, classical theory is test-oriented rather than item-oriented and thus does not make predictions about how an individual or group will perform on a particular item.
Reliability and Construct Validation

These limitations can be addressed in IRT (see Morizot, Ainsworth, & Reise, Chapter 24, this volume). Briefly put, IRT provides quantitative procedures to describe the relation of a particular item to the latent construct being measured in terms of difficulty and discrimination parameters. This information can be useful for item analysis and scale construction, permitting researchers to select items that best measure a particular level of a construct and to detect items biased for particular respondent groups. IRT is increasingly being applied to personality measures, such as self-esteem (Gray-Little et al., 1997) and romantic attachment (Fraley et al., 2000).

To summarize, in this section we focused on classical test theory approaches to reliability, the costs associated with low reliability and the practice of correcting for attenuation, specific types of reliability indices, and issues with coefficient alpha (e.g., test length, unidimensionality, and construct definitions). In our discussion, we mentioned such concepts as latent (or underlying) constructs, construct definitions, dimensionality, criterion variables, and discriminant relations, but did not discuss them systematically. These concepts are complex and go beyond the classical view of reliability, emphasizing that the meaning and interpretation of measurements is crucial to evaluating the quality of our measurements. Traditionally, issues of score meaning and interpretation are discussed under the heading of validity, to which we now turn.

Construct Validation

Measurements of psychological constructs, such as neuroticism, rejection sensitivity, or smiling, are fundamentally different from basic physical measurements (e.g., mass), which can often be based on concrete standards—such as the 1-kilogram chunk of platinum and iridium that standardizes the measurement of mass. Unfortunately, the behaviorist movement sparked a preoccupation with “gold standards” (or platinum-and-iridium standards) for psychological measures (e.g., Cureton, 1951) that lasted into the 1970s (see Kane, 2004). Eventually researchers came to recognize that for most psychological concepts there exists no single, objective, definitional criterion standard against which all other such measures can be compared.

In the absence of such criterion standards, personality psychologists have long been concerned with ways to conceptualize the validity of their measurement procedures. Although the first American Psychological Association committee on psychological tests distinguished initially among several “types” of validity, Cronbach and Meehl (1955) had already recognized that all validation of psychological measures is fundamentally concerned with what they called construct validity—evidence that scores on a particular measure can be interpreted as reflecting variation in a particular construct (i.e., an inferred characteristic) that has particular implications for human behavior, emotion, and cognition.

Process and Evidence

The idea of construct validity has been elaborated upon over time by such investigators as Loevinger (1957), Cronbach (1988), and Messick (1995), and it is now generally recognized as the central concern in psychological measurement (see also Braun, Jackson, & Wiley, 2002; Kane, 2004). The 1999 edition of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) emphasizes that “validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use” (p. 9). Yet construct validity continues to strike many of us, from graduate students to senior professors, as a rather nebulous or “amorphous” concept (Briggs, 2004), perhaps because there is no such thing as “the construct validity coefficient,” no single statistic that researchers can point to as proof that their measure is valid. Because of this, it may be easier to think of construct validity as a process (i.e., the steps that one would follow to test whether a particular interpretation of a particular measure is valid) than as a property (i.e., the specific thing that a measurement interpretation must have in order to be valid).

Keeping with this emphasis on construct validity as a process rather than a property, Smith (2005b) articulated four key steps in the validation process. First, a definition of the theoretical construct to be measured is proposed. Second, a theory, of which the construct in
question is a part, is translated into hypotheses about how a valid measure of the construct would be expected to act. Third, research designs appropriate for testing these hypotheses are formulated. Fourth, data are collected using these designs and observations based on these data are compared to predictions. (For a similar, process-oriented approach based on definition and evidence, see Kane, 2004.)

A fifth step, revision of the theory, construct, or measure (and repetition of steps one through four), highlights the idea that the construct validation process is a basic form of theory testing: “To validate a measure of a construct is to validate a theory” (Smith, 2005a, p. 413). As with any other theory or model, the validity of the particular score interpretation can never be fully established but is always evolving to form an ever-growing “nomological network” of validity-supporting relations (Wiggins, 1973).

Given that multiple pieces of evidence are needed to cumulatively support the hypothesized construct, it is often difficult to quickly summarize the available validity evidence. For example, Snyder (1987) wrote an entire book to summarize the validity argument (Cronbach, 1988) for his self-monitoring construct, drawing on everything that had been learned about this construct in more than 15 years of empirical research and construct development. More recently, meta-analytic techniques have proven useful to make such data summaries more manageable and objective (Schmidt, Hunter, Pearlman, & Hirsch, 1985). Westen and Rosenthal (2003) proposed two heuristic indices to operationalize construct validity in terms of the relative fit of observations to hypotheses, thus addressing the fourth step in Smith’s (2005b) process model. Nonetheless, attempts to quantify construct validity remain controversial (cf. Smith, 2005a, 2005b; Westen & Rosenthal, 2005).

Another way to elaborate the notion of construct validity in personality psychology is to consider the kinds of evidence that personality psychologists typically seek as part of the construct validation process (cf. Messick, 1989, 1995). Here we focus on five major forms of

<table>
<thead>
<tr>
<th>Forms of evidence for construct validity</th>
<th>Examples of study designs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Generalizability: Evidence that score properties and interpretations generalize across population groups, settings, and tasks</td>
<td>Reliability and replication: Test whether score properties are consistent across occasions (i.e., retest reliability), samples, and measurement methods (e.g., self-report and peer report)</td>
</tr>
<tr>
<td>2. Content validity: Evidence of content relevance, representativeness, and technical quality of items</td>
<td>Expert judgments and review: Test whether experts agree that items are relevant and represent the construct domain; use ratings to assess item characteristics, such as comprehensibility and clarity</td>
</tr>
<tr>
<td>3. Structural validity: Evidence that the internal structure of the measure reflects the internal structure of the construct domain</td>
<td>Exploratory or confirmatory factor analysis: Test whether the factor structure of the measure matches the hypothesized structure of the construct</td>
</tr>
<tr>
<td>4. External validity: Evidence that the measure relates to other measures and to non-test criteria in theoretically expected ways</td>
<td>Criterion correlation: Test whether measurement scores correlate with relevant criteria (e.g., membership in a criterion group)</td>
</tr>
<tr>
<td>5. Substantive validity: Evidence that measurement scores meaningfully relate to theoretically postulated domain processes</td>
<td>Multitrait–multimethod matrix: Test whether different measures of the same construct correlate more strongly than do measures of different constructs that use the same and different methods (e.g., instruments, data sources)</td>
</tr>
<tr>
<td>Mediation analysis: Test whether measurement scores mediate the relationship between an experimental manipulation and a behavioral outcome in an expected way</td>
<td></td>
</tr>
</tbody>
</table>
We emphasize at the outset that modest levels of generalization, and the boundaries of evidence that should be considered in the validation process, that particular kinds of evidence may be more or less important for supporting the validity of a particular measurement interpretation, and that these five kinds of evidence are not intended as mutually exclusive categories.

Evidence for Generalizability

Generalizability evidence is needed in a test validation program to demonstrate that score interpretations apply across tasks or contexts, times or occasions, and observers or raters (see Table 27.2). The inclusion of generalizability evidence here makes explicit that construct validation includes consideration of “error associated with the sampling of tasks, occasions, and scorers [that] underlie traditional reliability concerns” (Messick, 1995, p. 746). That is, the notion of generalizability encompasses traditional conceptions of both reliability and criterion validation; they may be considered on a continuum, differing only in how far generalizability claims can be extended (Thorndike, 1997). Traditional reliability studies provide relatively weak tests of generalizability, whereas studies of criterion validity provide stronger tests of generalizability.

For example, generalizing from a test score to another test developed with parallel procedures (e.g., a Form A and Form B of the same test) does increase our confidence in the test but does so only modestly (i.e., providing evidence of parallel-form equivalence). If we find we can also generalize to other times or occasions, our confidence is further strengthened, but not by quite as much as when we can show generalizability to other methods or even to nonexistent criteria related to the construct the test was intended to measure. Thus, generalizability can be thought of as similar to an onion—not because it smells bad, but because it involves layers. The inner layers represent relatively modest levels of generalization, and the outer layers represent farther-reaching generalizations to contexts that are more and more removed from the central core (i.e., dissimilar from the initial measurement operation).

The kind of validity evidence Messick (1989) considered under the generalizability rubric is crucial for establishing the limits or boundaries beyond which the interpretation of the measure cannot be extended. An issue of particular importance for personality researchers is the degree to which findings generalize from “convenience” samples, such as American college students, to groups that are less educated, older, or come from different ethnic or cultural backgrounds.

Evidence for Content Validity

A second form of validational evidence involves content validity; such evidence is provided most easily if the construct has been explicated theoretically in terms of specific aspects that exhaust the content domain to be covered by the construct. Common problems involve underrepresenting an important aspect of the construct definition in the item pool and overrepresenting another one. An obvious example is the multiple-choice exams we often construct to measure student performance in our classes; if the exam questions do not sample fairly from the relevant textbook and lecture material, we cannot claim that the exam validly represented what students were supposed to learn (i.e., the course content).

Arguments about content validity arise not only between professors and students, but also in research. The Self-Monitoring Scale (see Snyder, 1974) is a good example because it began with a set of 25 rationally derived items; when evidence later accumulated regarding the structure and external correlates of these items, Snyder (e.g., 1987) made revisions to both the construct and the scale, excluding a number of items measuring other-directed self-presentation. As a result, behavioral variability and attitude–behavior inconsistency were represented to a lesser extent in the revised scale. Because all items measuring public performing skills were retained, the construct definition in the new scale shifted toward a conceptually unrelated construct, extraversion (John, Cheek, & Klohnen, 1996). This example shows that discriminant aspects are also important in content validation: To the extent that the items measure aspects not included in the construct definition, the measure would be contaminated by construct-irrelevant variance. For example, when validating scales to measure coping or emotion regulation, the item content on such scales should not assess variance that must be attributed to distinct constructs, such as psychological adjustment or social outcomes that...
are theoretically postulated to be direct consequences of the regulatory processes the scales are intended to assess (see, e.g., John & Gross, 2004, 2007).

To address questions about content validity, researchers may use a number of validation procedures (see also Smith & McCarthy, 1995). Researchers might ask expert judges to review the match between item representation and construct domain specification, and these conceptual-theoretical judgments can then be used to add or delete items. For example, Jay and John (2004) adopted the Diagnostic and Statistical Manual of Mental Disorders (4th ed.) (DSM-IV) symptom list for major depression as an explicit construct definition for their California Psychological Inventory (CFI)-based Depressive Symptom (DS) scale. Advanced graduate students in clinical psychology then provided expert judgments classifying all the proposed DS items as well as the items of several other depression self-report scales according to the DSM symptoms. Moreover, in an effort to address discriminant validity early in the scale construction process, the judges were also given the choice to classify items as more relevant to anxiety than to depression, reasoning that depression needs to be conceptually differentiated from anxiety and therefore anxiety items should not appear on a depressive symptom scale. In this way, Jay and John (2004) were able to (1) focus their scale on item content uniquely related to depression rather than anxiety symptoms, (2) examine how comprehensively their DS item set represented the intended construct (e.g., they found that only one DSM symptom cluster, suicidal ideation, was not represented), and (3) compare the construct representation of the DS items with those of other, commonly used depression scales. This theoretically based approach, when applied to questionnaire construction, has become known as the rational-intuitive approach; it has been widely used by personality and social psychologists focused on measuring theoretically postulated constructs (e.g., Burisch, 1986). Probably the most explicitly rational approach to construct definition in personality psychology is Buss and Craik’s (1983) act frequency approach: Selection of act items was based not on an abstract theoretical definition of each trait construct but on folk wisdom, captured in terms of college students’ aggregated judgments of the prototypicality (or relevance) of a large number of acts for a particular trait (Buss & Craik, 1983).

Content validity can also be considered in the context of the quality and adequacy of formal or technical aspects of items. In the domain of self-reports and questionnaires, it is important to recognize that the researcher is trying to communicate accurately and efficiently with the research participants, and thus formal item characteristics, such as the clarity of wording, easy comprehensibility, low ambiguity, and so on, are crucial linguistic and pragmatic concerns in the design of items (e.g., Angleitner, John, & Lohr, 1986).

Evidence for Structural Validity

Structural validity requires evidence that the correlational (or factor) structure of the items on the measure is consistent with the hypothesized internal structure of the construct domain. We noted the issue of multidimensionality in the section on reliability, pointing out that coefficient alpha does not allow inferences about the dimensionality of a measure. The structure underlying a measure or scale is not an aspect of reliability; rather, it is central to the interpretation of the resulting scores and thus needs to be addressed as part of the construct validation program. Researchers have used both exploratory and confirmatory factor analysis for this purpose; we return to this important issue below in the context of evaluating measurement with structural equations models.

Evidence for External Validity: Convergent and Discriminant Aspects

External validity has been at the core of what most personality psychologists think validity is all about: How well does a test predict conceptually relevant behaviors, outcomes, or criteria? Wiggins (1973, p. 406) argued that prediction “is the sine qua non of personality assessment,” and Dawes and Smith (1985, p. 512) suggested that “the basis of all measurement is empirical prediction.” Obviously, it makes sense that a test or scale should predict construct-relevant criteria. It is less apparent that we also need to show that the test does not predict conceptually unrelated criteria. In other words, a full demonstration of external aspects of construct validation requires a demonstration of both what the test measures and what it does not measure.
Reliability and Construct Validation

Predicting Criterion Group Membership and Nontest (External) Criterion Variables

One long-popular method for demonstrating the external validity of a measure was to test whether the measure can successfully distinguish between criterion groups—groups that are presumed to differ substantially in their mean levels of the construct to be measured. Historically, the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1943) and CPI (Gough, 1957) were the first personality inventories developed according to the criterion (or contrast) group approach. For example, items were selected for the MMPI depression scale if they could discriminate patients hospitalized with a diagnosis of major depression from nonpsychiatric control subjects. Gough (1957) selected items for his subsequent achievement via conformance scale if they predicted grade point average (GPA) in high school (assumed to reflect conventional achievement requiring rule following) and for his achievement via independence scale if they predicted GPA in college (assumed to reflect more autonomous pursuit of achievement goals and interests). More recently, Cacioppo and Petty (1982) developed the need for cognition scale to measure individual differences in the preference and enjoyment of effortful thinking. As part of their construct validation program, they conducted a study contrasting college professors (assumed to need cognition) and assembly line workers (assumed not to need cognition). Consistent with the interpretation of their measure as reflecting individual differences in need for cognition, the mean score of the professors was much higher than the mean score of the assembly line workers. Gosling, Kwan, and John (2003) validated owners’ judgments of their dogs’ Big Five personality traits by showing that these judgments predicted relevant behavior in a dog park, as rated by strangers who interacted with the dogs for an hour.

A critical issue with the use of such external criteria is the “gold standard” problem mentioned earlier—that the convergent and discriminant construct validity of the criterion itself is typically not well established. For example, patients with a diagnosis of major depression may be comorbid with other disorders (e.g., anxiety) or may have been hospitalized for construct-irrelevant reasons (e.g., depressed individuals lacking social or financial support are more likely to be hospitalized), just as college professors likely differ from assembly workers in more ways than just their personal need for cognition. In recognition of this problem, Gosling, Kwan, and John (2003) tried to rule out potential confounds, such as that the observers’ behavior ratings of the dogs in the park were not based simply on appearance and breed stereotypes that may be shared by both owners and strangers.

Multitrait–Multimethod Matrix

Campbell and Fiske (1959) introduced the terms convergent and discriminant to distinguish demonstrations of what a test measures from demonstrations of what it does not measure. The convergent validity of a self-report scale of need for cognition could be assessed by correlating the scale with independently obtained peer ratings of the subject’s need for cognition and with frequency of effortful thinking measured by “beeping” the subject several times during the day. Discriminant validity could be assessed by correlating the self-report scale with peer ratings of extraversion and a beeper-based measure of social and sports activities. Campbell and Fiske were the first to formalize these ideas of convergent and discriminant validity into a single systematic design that crosses multiple traits or constructs (e.g., need for cognition and extraversion) with multiple methods (e.g., self-report, peer ratings, and beeper methodology). They called this design a multitrait–multimethod (MTMM) matrix, and the logic of the MTMM is both intuitive and compelling.

What would we expect for our need for cognition example? Certainly, we would expect sizable convergent validity correlations between the need for cognition measures across the three methods (self-report, peer report, beeper); because these correlations involve the same trait but different methods, Campbell and Fiske (1959) called them monotrait–heteromethod coefficients. Moreover, given that need for cognition is theoretically unrelated to extraversion, we would expect small discriminant correlations between the need for cognition measures and the extraversion measures. This condition should hold even if both traits are measured with the same method, leading to so-called heterotrait–monomethod correlations. Certainly, we want each of the convergent correlations to be substantially higher than the discriminant correlations involving the same trait. And finally, the same patterns of intercorre-
lations between the constructs should emerge, regardless of the method used; in other words, the relations between the constructs should generalize across methods.

**Method Variance**

An important recognition inherent in the MTMM is that we can never measure a trait or construct by itself; rather, we measure the trait intertwined with the method used: "Each measure is a trait–method unit in which the observed variance is a combined function of variance due to the construct being measured and the method used to measure that construct" (Rezmovic & Rezmovic, 1981, p. 61). The design of the MTMM is so useful because it allows us to estimate variance in our scores that is due to method effects—that is, errors systematically related to our measurement methods and thus conceptually quite different from the notion of random error in classical test theory. These errors are systematic because they reflect the influence of unintended constructs on scores, that is, unwanted variance—something we did not wish to measure but that is confounding our measurement (Ozer, 1989).

Method variance is indicated when two constructs measured with the same method (e.g., self-reported attitudes and self-reported behavior) correlate more highly than when the same constructs are measured with different methods (e.g., self-reported attitudes and behavior coded from videotape). Response styles, such as acquiescence, may contribute to method variance when the same respondent completes more than one measure (Soto, John, Gosling, & Potter, 2006). Another example involves positivity bias in self-perceptions, which some researchers view as psychologically healthy (Taylor & Brown, 1988). However, if positivity bias is measured with self-reports and the measure of psychological health is a self-report measure of self-esteem, then a positive intercorrelation between these measures may not represent a valid hypothesis about two constructs (positivity bias and psychological health), but shared self-report method variance associated with narcissism (John & Robins, 1994); that is, individuals who see themselves too positively may be narcissistic and also rate their self-esteem too highly. Discriminant validity evidence is needed to rule out this alternative hypothesis, and the construct validity of the positivity bias measure would be strength-en ed considerably if psychological health were measured with a method other than self-report, such as ratings by clinically trained observers (Jay & John, 2004).

**LOTS: Multiple Sources of Data**

Beginning with Cattell (1957, 1972), psychologists have tried to classify the many sources researchers can use to collect data into a few broad categories. Because each data source has unique strengths and limitations, the construct validation approach emphasizes that we should collect data from lots of different sources, and so the acronym LOTS has particular appeal (Block & Block, 1980; see also Craik, 1986).

LOTS data refer to life event data that can be obtained fairly objectively from an individual’s life history or life record, such as graduating from college, getting married or divorced, moving, socioeconomic status, memberships in clubs and organizations, and so on. Examples of particularly ingenious measures derived from LOTS data are counts of bottles and cans in garbage containers to measure alcohol consumption (Webb, Campbell, Schwartz, Sechrest, & Grove, 1981), police records of arrests and convictions to measure antisocial behavior (Caspi et al., 2003), and the use of occupational, marital, and family data to score the number of social roles occupied by an individual (Helson & Soto, 2005).

LOTS data refer to observational data, ranging from observations of very specific aspects of behavior to more global ratings (see Bakeman, 2000; Kerr, Aronoff, & Messé, 2000). Examples are careful and systematic observations recorded by human judges, such as in a particular laboratory setting or carefully defined situation; behavior coded or rated from photos or videos; and, broader still, reports from knowledgeable informants, such as peers, roommates, spouses, teachers, and interviewers that may aggregate information across a broad range of relevant situations in the individual’s daily life. O data obtained through unobtrusive observations or coded later from videotape can be particularly useful to make inferences about the individual’s attitudes, prejudices, preferences, emotions, and other attributes of interest to social scientists. Harker and Keltner (2001) used ratings of emotional expressions in women’s college yearbook photos to predict marital and well-being outcomes 30 years later. Gross and Levenson (1993) used frequency of
Reliability and Construct Validation

...as an index of distress. Fraley and Shaver (1998) observed and coded how different romantic couples behaved as they were saying good-bye to one another at an airport and found that separation behavior was related to adult attachment style. Another nice illustration is a study that recorded seating position relative to an outgroup member to measure ethnocentrism (Macrae, Bodenhausen, Milne, & Jetten, 1994).

T data refer to information from test situations that provide standardized measures of performance, motivation, or achievement, and from experimental procedures that have clear and objective rules for scoring performance. A timed intelligence test is the most obvious example; other examples include assessments of the length of time an individual persists on a puzzle or delays gratification in a standardized situation (Ayduk et al., 2000). Reaction times are frequently used in studies of social cognition, providing another kind of objective measure of an aspect of performance. Recently, the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), which uses reaction-time comparisons to infer cognitive associations, has become a popular method of assessing implicit aspects of the self-concept. Greenwald and Farnham (2000) provided evidence for the external and discriminant validity of IAT measures of self-esteem and masculinity-femininity.

Finally, S data refer to self-reports. S data may take various forms. Global self-ratings of general characteristics and true-false responses to questionnaire items have been used most frequently. However, self-reports are also studied in detailed interviews (see Bartholomew, Henderson, & Marcia, 2000), in narratives and life stories (see Smith, 2000), and in survey research (Visser, Kroesnick, & Lavrakas, 2000). Daily experience sampling procedures (see Reis & Gable, 2000) can provide very specific and detailed self-reports of moment-to-moment functioning in particular situations.

The logic underlying S data is that individuals are in a good position to report about their psychological processes and characteristics—unlike an outside observer, they have access to their private thoughts and experiences and they can observe themselves over time and across situations. However, the validity of self-reports depends on the ability and willingness of individuals to provide valid reports, and self-reports may be influenced by various constructs other than the intended one. Systematic errors include, most obviously, individual differences in response or rating scale use, such as acquiescence (see McCrae, Herbst, & Costa, 2001; Soto, John, Gosling, & Potter, 2006; Visser et al., 2000) and response extremeness (Hamilton, 1968). Another potential source of error is reconstruction bias, in which individuals' global or retrospective ratings of emotions and behaviors differ substantially from their real-time or "online" ratings (Scollon, Diener, Oishi, & Biswas-Diener, 2004).

Moreover, some theorists have argued that self-reports are of limited usefulness because they may be biased by social desirability response tendencies. Two kinds of desirability biases have been studied extensively (for a review, see Paulhus, 2002; also Paulhus & Vazire, Chapter 13, this volume). Impression management refers to deliberate attempts to misrepresent one's characteristics (e.g., "faking good") whereas self-deceptive enhancement reflects honestly held but unrealistic self-views. Impression management appears to have little effect in research contexts where individuals participate anonymously and are not motivated to present themselves in a positive light; self-deception is not simply a response style but related to substantive personality characteristics, such as narcissism (Paulhus & John, 1998).

Fortunately, although personality psychologists still use self-report questionnaires and inventories most frequently, other methods are available and used (Craik, 1986; Craik, Chapter 12, this volume; Robins, Tracy, & Sherman, Chapter 37, this volume). Thus, measures based on L, O, and T data can help evaluate and provide evidence for the validity of more easily and commonly obtained self-report measures tapping the same construct. Unfortunately, research using multiple methods to measure the same construct has not been very frequent. Overall, multimethod designs have been underused in construct validation efforts. Researchers seem more likely to talk about the MTMM approach than to go to the trouble of actually using it.

There is an extensive and useful methodological literature on the MTMM, which began in the mid-1970s when SEM became available and provided powerful analytical tools to estimate separate trait and method factors (e.g., Kenny, 1976; Schwarz, 1986; Wegener...
A number of excellent reviews and overviews are also available. For example, Judl and McClelland (1998) describe a series of examples that illustrate Campbell and Fiske's (1959) original principles of convergent and discriminant validation as well as the application of SEM techniques to estimate separate trait and method effects. For specific issues in fitting SEM models, see Kenny and Kashy (1992) and Marsh and Grayson (1995). Hypothetical data may be found in West and Finch (1997), who illustrate three scenarios: (1) convergent and discriminant validity with minimal method effects, (2) strong method effects, and (3) effects of unreliability and lack of discriminant validity. John and Srivastava (1999) modeled trait and instrument effects with data for three commonly used Big Five instruments.

Evidence for Substantive Validity

The final form of validational evidence in Table 27.5 involves substantive validity. Substantive validation studies make use of substantive theories and process models to further support the interpretation of the test scores. The strongest evidence for substantive validity comes from studies that use experimental manipulations that directly vary the processes in question. For example, Petty and Cacioppo (1986) showed that the process of attitude change was mediated by need for cognition. Individuals scoring high on the scale were influenced by careful examination of the arguments presented in a message, whereas those scoring low were more influenced by extraneous aspects of the context or message (e.g., the attractiveness of the source of the message). Another example is Paulhus, Bruce, and Trapnell's (1993) use of experimental data to examine an aspect of the substantive validity of two social desirability scales. When subjects were asked to intentionally present themselves in a favorable way (e.g., as they might during a job interview), the self-presentation scale showed the predicted increase over the standard "honest self-description" instruction, but the self-deception scale did not, just as one would expect for a scale designed to measure unrealistically positive self-views that the individual believes are true of him or her.

Substantive validity, then, is really about testing theoretically derived propositions about how the construct in question should function in particular kinds of contexts and how it should influence the individual's behavior, thoughts, or feelings. In that sense, studies of substantive validity are at the boundary between validational concerns and broader concerns with theory building and testing. The concept of substantive validity thus serves to illustrate the back and forth (dialectic) of theory and research. That is, when a study fails to show the effect predicted for a particular construct, it is unclear whether the problem involves a validity issue (i.e., the measure is not valid), or faulty theorizing (i.e., the theory is wrong), or both.

The consideration of substantive aspects of validity illustrates that ultimately measurement cannot be separated from theory, and a good theory is one that includes an account of the relevant measurement properties of its constructs. For example, a theory of emotion might distinguish among multiple emotion components, such as subjective experience, emotion-expressive behavior, and physiological response patterns (see, e.g., Gross, 1999), and specify how these components are most validly measured, such as emotion experience with particular kinds of self-report measures, emotion expressions with observer codings from video recordings of the individual, and physiological responding with objective tests. How exactly these three emotion components ought to be related to each other is foremost a theoretical issue but also involves substantive validity issues; for example, if a study were to show zero correlations between measures of sadness experience and measures of sadness expression, the theoretical notion of emotion as a unitary and coherent construct may have to be modified because method variance and systematic factors (e.g., display rules; individual differences in expressivity) might influence the coherence of emotion experience and expression for particular emotions and particular individuals (see, e.g., Gross, John, & Richards, 2000). Substantive validity, then, is the broadest of all five forms of validational evidence and, ultimately, indistinguishable from using theory-testing to build the nomological network for a construct.

Construct Validation: Summary and Conclusions

To summarize, in this section we reviewed five forms of evidence central to a program of construct validation (see Table 27.5). We consid-
Reliability and Construct Validation

Measurement Models in SEM: Convergent Validity, Discriminant Validity, and Random Error

Like all factor analytic procedures (Floyd & Widaman, 1995; Tinsley & Tinsley, 1987; see also Lee & Ashton, Chapter 25, this volume), CFA assumes that a large number of observations or items are a direct result (or expression) of a smaller number of latent sources (i.e., unobserved, hypothetical, or inferred constructs). However, CFA eliminates some of the arbitrary features often criticized in exploratory factor analysis (Gould, 1981; Sternberg, 1985). First, CFA techniques require the researcher to specify an explicit model (or several competing models) of how the observed (or measured) variables are related to the hypothesized latent factors. Second, CFA offers advanced statistical techniques that allow the researcher to test how well the a priori model fits the particular data; even more important, CFA permits comparative model testing to establish whether the a priori model fits the data better (or worse) than plausible alternative or competing models.

CFA models can also be displayed graphically, allowing us to effectively communicate the various assumptions of each model. Two examples are shown in Figure 27.3. Figure 27.3a shows a common-factor model in which a single underlying construct (neuroticism, shown as an ellipse at the top) is assumed to give rise to the correlations between all 12 items, or responses A1 to D6 (the observed variables, shown in squares). Following convention (Bentler, 1980), ellipses are used to represent latent variables, whereas squares represent measured (or manifest) variables; arrows with one head represent directed or regression parameters, whereas two-headed arrows (which are often omitted) represent covariance of undirected parameters. Note that each measured variable has two arrows leading to it. The arrow from the latent construct is a factor

Model Testing in Construct Validation and Scale Construction

The measurement model in structural equation modeling (SEM; Jöreskog & Sörbom, 1981; see also Bentler, 1980) is based on confirmatory factor analysis (CFA). Kline (2004), Loehlin (2004), McArdle (1996), and Bollen and Long (1993) have provided readable introductions. CFA is particularly promising because it provides a general analytic approach to assessing construct validity. As will become clear, convergent validity, discriminant validity, and random error can all be addressed within the same general framework. To illustrate these points, we return to our earlier numerical examples (Tables 27.3 and 27.4) and show how these data can be analyzed and understood using CFA-based measurement models.

Model Testing in Construct Validation and Scale Construction

The measurement model in structural equation modeling (SEM; Jöreskog & Sörbom, 1981; see also Bentler, 1980) is based on confirmatory factor analysis (CFA). Kline (2004), Loehlin (2004), McArdle (1996), and Bollen and Long (1993) have provided readable introductions. CFA is particularly promising because it provides a general analytic approach to assessing construct validity. As will become clear, convergent validity, discriminant validity, and random error can all be addressed within the same general framework. To illustrate these points, we return to our earlier numerical examples (Tables 27.3 and 27.4) and show how these data can be analyzed and understood using CFA-based measurement models.
FIGURE 27.3. Two confirmatory factor analysis models of 12 NEO-PI-R neuroticism items (6 from the anxiety (A) and 6 from the depression (D) facet scale (see Table 27.4 and Figures 27.2 and 27.4). In each panel, the top row of values represents estimates of standardized regression coefficients (i.e., factor loadings in correlational metric). The bottom row of values provides estimates of error variances (i.e., uniquenesses expressed in variance terms). Panel (a) shows parameter estimates for the general neuroticism factor model (Model 1 in Table 27.6); panel (b) shows parameter estimates for the best-fitting two-correlated-factors model with item A4 ("I often feel tense and jittery") allowed to load on both the anxiety and depression factors (Model 4 in Table 27.6).
loading $L_m$ that represents the strength of the effect that the latent construct has on each observed variable. The other arrow involves another latent variable for each observed variable—these are unique factor scores ($e_m$) that represent the unique or residual variance ($U^2$) remaining in each observed variable.

Conceptually, this model captures a rather strong structural hypothesis, namely, that the 12 observed variables covary only because they all measure the same underlying construct, neuroticism. In other words, we hypothesize that the only thing the items have in common is this one latent construct, and all remaining or residual item variance is idiosyncratic to each item and thus unshared. This structural model provides a new perspective on how to define two important terms we have used in this chapter: the convergent validity of the item and random error. In particular, the loading of an item on the construct of interest represents the convergent validity of the item, whereas its unique variance represents random error. However, in this simple measurement model, we cannot address discriminant validity.

Compare the measurement model in Figure 27.3a to the one in Figure 27.3b, which postulates two factors (anxiety and depression) influencing responses to the same 12 items. Here we are hypothesizing two distinct constructs, rather than one. This model incorporates another condition, known as simple structure. The convergent validity loadings (represented by arrows from the latent constructs to the observed items) indicate that the first 5 items are influenced by the first construct but not the second construct, whereas the last 6 items are influenced only by the second construct and not the first. In other words, 11 of these items (all except item A4) can be uniquely assigned to only one construct, thus greatly simplifying the measurement model.

With two constructs in the measurement model, we can also address issues of discriminant validity. Whereas an item's loading on the construct of interest represents convergent validity and its unique variance random error, its loading on a construct other than the intended one speaks to its discriminant validity. For item A4, our earlier correlational analyses had suggested discriminant validity problems (see Table 27.4), and we therefore examined one model (Model 4, which is shown in Figure 27.3b) that allowed item A4 to have both a convergent loading (on its intended anxiety factor) and a discriminant loading (on the depression factor).

Note that this model includes an arrow between the two constructs, indicating a correlation (or covariance); the two constructs are not independent (orthogonal) but related (oblique). This correlation tells us about discriminant validity at the level of the constructs. If the correlation is very high (e.g., .90), we would worry that the two constructs are not sufficiently distinct and that we really have only one construct. If the correlation is quite low (e.g., .10), we would be reassured that the two concepts show good discriminant validity with respect to each other. There is another possibility here, namely, that the two constructs are substantially correlated (e.g., .70) because they are both facets, or components, of a broader, superordinate construct that includes them both, which, of course, is the hierarchical model of the NEO-PI-R from which this example was taken. Note that here we are addressing issues that involve questions about the dimensionality and internal structure of the constructs being measured. We discussed these issues earlier in the section on reliability, especially coefficient alpha, but, as we argued in the section on validity, dimensionality issues should be considered part of the construct validation program (see Table 27.5) because they concern the structural validity of the interpretation of our measures.

Structural Validity Examined with SEM

Structural validity issues resurface with great regularity in the personality literature. Some of the most popular constructs have endured protracted debates focused on their structural validity: self-monitoring, attributional style, hardness, Type A coronary-prone behavior pattern, and, most recently, need for closure (e.g., Hull, Lehn, & Tedlie, 1991; Neuberg, Judd, & West, 1997). Part of the problem is that many of these constructs, and the scales designed to measure them, were initially assumed to be unidimensional, but later evidence challenged those initial assumptions. It is therefore instructive to consider how SEM approaches can help address the underlying issues and to provide some numerical examples to illustrate the issues.
Testing the Unidimensionality of Scales: What Coefficient Alpha Reliability Cannot Do

As we noted earlier, coefficient alpha does not address whether a scale is unidimensional. For this purpose, factor analyses are needed; CFA provides the most rigorous approach because it can test how well the interitem correlation matrix for a particular scale fits a single-factor, rather than multifactor, model. In other words, how well can the loadings on a single factor reproduce the correlation matrix actually observed?

Consider again the three scales for which we presented interitem correlation data in Table 27.3; all had alphas of .87. What do CFAs of these correlation matrices tell us about their dimensionality? One-factor models perfectly fit the data pattern for both Scales A and B, just as expected for these unidimensional scales. CFA also estimates factor loadings for the items, providing an index of content saturation. For Scale A (which had nine items all intercorrelating .42), the items all had the same factor loading of .648 (i.e., the square root of .42, which was the size of the interitem correlations in this example). For Scale B (six items all intercorrelating .52), the factor loadings were all .721; this slightly higher value reflects that the interitem correlations (and thus content saturation) were slightly higher for Scale B than for Scale A.

In contrast, for Scale C (which had heterogeneous interitem correlations of .40 and .70, averaging .52) the fit of the one-factor model was unacceptable (for N = 200, CFI = .726, and root mean square error of approximation [RMSEA] = .263). Note, however, that the item loadings were all .721 (i.e., the square root of .52, which was the mean of the interitem correlations), the same as for the truly unidimensional Scale B. As expected, the two-factor model fit Scale C better than did the one-factor model, and perfect fit was obtained when we allowed the factors to correlate. Reflecting their .70 correlations with each other, Scale C items 1, 2, and 3 loaded .837 on factor 1 and 0 on factor 2, whereas items 4, 5, and 6 loaded 0 on factor 1 and .837 on factor 2. The interitem correlation of .40 across the two subsets of items was reflected in an estimated correlation of .571 between the two latent factors. These results highlight that the error (or unreliability) present in an item is separate from the issue of multidimensionality. In these CFA measurement models, the item loadings represent how much of the item variance is shared across items (and is thus generalizable). Error is captured by the residual item variance (i.e., 1 minus the squared loading) indicating how much variance is unique to that item; the proportion of shared to total item variance is often referred to as content saturation. Dimensionality, however, is captured by the relative fit of the one-factor model versus multiple-factor models. Comparing Scales A and C in Table 27.3, the longer Scale A is clearly more unidimensional than C, yet its items do not show greater content saturation (i.e., higher factor loadings and lower error terms). In other words, unidimensionality does not imply lower levels of measurement error (i.e., unreliability), and vice versa.

Comparative Tests of Measurement Models: Testing Alternative Models for the Neuroticism Example

The general model is shown in Figure 27.2, along with the mean interitem correlations and alpha reliability coefficients we obtained when we applied the traditional canon of internal consistency analysis to these scales: a 12-item superordinate neuroticism scale with 6-item anxiety and depression facets. The traditional method in the analysis of structural validity is exploratory factor analysis. When applied to this example (see the interitem correlation matrix given in Table 27.4), we expected evidence for both a general neuroticism factor (the first unrotated principal component) and two rotated factors representing the anxiety and depression items, respectively. Indeed, principal components analyses resulted in eigenvalues that made it difficult to decide between the one- and two-factor solutions: The first unrotated component accounted for 39% of the total variance, almost four times the size of the second component, which accounted for only 10.8%. After varimax rotation, however, the two factors were almost identical in size, accounting for 26% and 23% of the variance.

The loadings for the two rotated factors are shown in Figure 27.4, with the x-axis representing the anxiety factor and the y-axis the depression factor. Overall, there was some evidence of simple structure, in that the items tended to cluster close to their anticipated fac-
FIGURE 27.4. Loading plot for exploratory factor (principal components) analysis of 12 NEO-PI-R neuroticism items, with 6 items each from the anxiety (A) and depression (D) facet scales (see Table 27.4 and Figure 27.2). N = 649 college students. An "(r)" following the item number indicates items to be reverse-keyed (i.e., items measuring low anxiety or low depression). The dashed line indicates the first unrotated component representing the general neuroticism factor. Note how the loadings for item A4 ("I often feel tense and jittery") place it about halfway between the anxiety and depression factor axes and close to the general neuroticism factor, suggesting that this item is a better indicator of general distress (Watson et al., 1995) shared by both anxiety and depression than a unique (or primary) indicator of anxiety.

that includes the low-anxiety and low-depression (reverse-keyed) items.

The dashed line represents the location of the first unrotated component, that is, the general neuroticism factor that accounted for almost 40% of the item variance; all 7 true-keyed items had substantial positive loadings on it, and all 5 false-keyed items had substan-
tial negative loadings. This factor clearly captures the positive correlation ($r = .58$) between the two item sets. Note how the loadings for item A4 ("I often feel tense and jittery") place it about halfway between the orthogonal anxiety and depression factor axes and rather close to the general neuroticism factor; this item may be a better indicator of general distress (Watson et al., 1995) shared by both anxiety and depression than a unique indicator of anxiety.

These exploratory factor analyses leave us with some alternative hypotheses that we can test formally using CFA. The CFA results are summarized briefly in Table 27.6. We begin with the one-factor model because it is the simplest or "compact model" (Judd et al., 1995). Because the models are all nested, we can statistically compare them with each other, testing the relative merits of more complex (i.e., full or augmented) models later. Without going into detail, the model-comparison results show that we can clearly reject Model 1 (one general neuroticism factor only) and Model 2 (two uncorrelated anxiety and depression factors), as both had substantially and significantly higher $\chi^2$ values than Model 3, which defines anxiety and depression as two distinct but correlated factors. Model 4, shown in Figure 27.3b, also takes into account the discriminant validity problems of item A4 by allowing it to load on the depression factor; this model fits significantly better than the simpler Model 3, though the change in $\chi^2$ was not large in size. These conclusions were also consistent with a wide variety of absolute and relative fit indices, like the Goodness-of-Fit Index (GFI) and the Comparative Fit Index (CFI). Table 27.6 also presents the estimated correlation between the two latent factors, which was .71 in Model 3. This value was, as expected, higher than the simple observed correlation of .58 between the unit-weighted scales and just a tad below the estimate of .74 for the correlation corrected for attenuation due to unreliability.4

More Complex Models
Including External Validity

In a fully developed construct validation program, of course, we would not stop here. Next, one might begin studies of external validity, modeling the relations of these two correlated CFA-based constructs with other measures of anxiety and depression, preferably drawn from other data sources, such as interview-based judgments by clinical psychologists (Jay & John, 2004). Using an MTMM design to address external validity, we would gather evidence about both convergent validity (e.g., self-reported anxiety with measures of anxiety drawn from another data source) and discriminant validity (e.g., self-reported anxiety with measures of depression drawn from another data source). Again, we would use SEM procedures for these additional validation steps, because one can model the measurement structure we have discussed so far, along with a predictive (or convergent) validity relation. Note that this model addresses the criterion problem that seemed so intractable in the early treatments of validity. The criterion itself is not treated as a "gold standard" but is modeled as a construct that must also be measured with fallible observed indicator variables. We should note that the models used to represent trait and method effects in MTMM matrices are considerably more complex than the simple models considered here. For example, McArdle (1996, Fig. 2) provides an elegant model for a more complete representation of the construct validation program.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$\Delta\chi^2$</th>
<th>GFI</th>
<th>CFI</th>
<th>Interfactor correlation</th>
<th>Mean loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. One general factor: neuroticism</td>
<td>504.5</td>
<td>54</td>
<td>212.1*</td>
<td>.87</td>
<td>.81</td>
<td>N/A</td>
<td>.57</td>
</tr>
<tr>
<td>2. Two uncorrelated factors: anxiety and depression</td>
<td>347.5</td>
<td>54</td>
<td>255.1*</td>
<td>.89</td>
<td>.79</td>
<td>.00</td>
<td>.62</td>
</tr>
<tr>
<td>3. Two correlated factors: anxiety and depression</td>
<td>292.4</td>
<td>53</td>
<td>N/A</td>
<td>.92</td>
<td>.90</td>
<td>.71</td>
<td>.62</td>
</tr>
<tr>
<td>4. Two correlated factors, plus one cross-loading item</td>
<td>271.9</td>
<td>52</td>
<td>20.5*</td>
<td>.93</td>
<td>.91</td>
<td>.67</td>
<td>.58</td>
</tr>
</tbody>
</table>

Note. $\Delta\chi^2$ compared with Model 3; GFI, Goodness-of-Fit Index (Jöreskog & Sörbom, 1981); CFI, Comparative Fit Index (Bentler, 1980); mean loading, mean of the standardized loading of each item on its factor(s).

* $p < .05$. 
Implications of Construct Validation for Scale Construction

So far, we have discussed construct validation as if the measure to be validated already existed. However, construct validation issues are central not only during the evaluation of existing measures but also during each stage of their development (see Simms & Watson, Chapter 14, this volume). Most modern scale construction efforts have adopted, implicitly or explicitly, many of the features of the construct validation program discussed in this chapter. In fact, much of our presentation here has spelled out the kinds of issues that researchers constructing a new measure must consider. There is no simple formula, but the integrated conception of construct validity and the various validation procedures summarized in Table 27.5 provide a blueprint for the kinds of evidence to be gathered and procedures to be followed.

Questionnaire construction, like measurement more generally, involves theory building and thus requires an iterative process. It begins with (1) generating hypotheses, (2) building a model and plausible alternatives, (3) generating items, using construct definitions, generalizability facets, and content validation procedures as guides (for information about item and response formats, see Visser et al., 2000), (4) gathering and analyzing data, (5) confirming and disconfirming the initial models, (6) generating alternative hypotheses leading to (7) improved models, (8) additional and more content-valid items, (9) more data gathering, and so on. The cycle continues, until a working model has been established that is "good enough"—one that the investigator can live with, for now, given the constraints and limits of real-life research. In other words, scale construction and construct validation go hand in hand, one cannot be separated from the other, and both fundamentally involve theory-building and theory-testing efforts.

Some Final Thoughts

In this chapter we have tried to strike a balance between description and prescription, between "what is" and "what should be" the practice of measurement and construct validation in personality research. We reviewed the traditional reliability coefficients but urged the reader to think instead about facets of generalizability such as time, items, and observers. We railed against some of our pet peeves, such as overreliance on coefficient alpha, articulating its limitations and arguing for a more nuanced understanding of this ubiquitous index. We advocated for a more process-oriented conception of construct validity, suggesting that the validation process deserves the same thoughtful consideration as any other form of theory testing. We illustrated, briefly, the power of no-longer new SEM techniques to help model measurement error, as well as convergent and discriminant validity.

This chapter has noted some shortcomings of current measurement conventions, practices that ought to be changed. Nonetheless, we are upbeat about the future. Specifically, over the past years we have become persuaded by the logic of comparative model testing; we now see it as the best strategy for evaluating and improving our measurement procedures. We are confident that as a new generation of personality researchers "grows up" using model comparison strategies, and as more of the old dogs among us learn this new trick, comparative model testing will continue to spread and help improve the validity of our measures. And because valid measures are a necessary precondition for good research, everything that we do as scientists comes back, in the end, to the importance of being valid.

Acknowledgments

The preparation of this chapter was supported, in part, by a grant from the Retirement Research Foundation and by a Faculty Research Award from the University of California-Berkeley. We are indebted to Monique Thompson, Josh Eng, and Richard W. Robins for their helpful comments on earlier versions.

Notes

1. Table 27.2 shows that both Pearson and intraclass correlations can be used to index retest stability. Pearson correlations reflect changes only in the relative standing of participants from one time to the other, which is typically the prime concern in research on individual differences. When changes in mean levels or variances are of interest too, then the intraclass correlation is the appropriate index.

2. In one context, this internal consistency concep-
tion does not apply. In most psychological measurement, the indicators of a construct are seen as effects caused by the construct; for example, individuals endorse items about liking noisy parties because of underlying individual differences in extraversion. However, as Bollen (1984) noted, constructs such as socioeconomic status (SES) are different. SES indicators, such as education and income, cause changes in SES, rather than SES causing changes in education or income. In these cases of “cause indicators,” the indicators are not necessarily correlated and the internal consistency conception does not apply.

3. As an additional, distinct form of validity evidence, Messick (1989) included what he called consequential validity, which addresses the personal and societal consequences of interpreting and using a particular measure in a particular way (e.g., using an ability test to decide on school admissions). It requires the test user to confront issues of test bias and fairness and is of central importance when psychological measures are used to make important decisions about individuals. This type of validity evidence is generally more relevant in applied research and in educational and employment settings than in basic personality research, in which scale scores have little, if any, consequence for the research participant.

4. The mean loadings in Table 27.6 indicate that loadings were substantial in all models, with the general-factor model falling just below the two-factor models. Figure 27.3b shows the final parameter estimates for Model 4; consistent with the discriminant-validity problems apparent in Table 27.4 and Figure 27.4, item A4 had significant loadings on both the anxiety and depression latent factors.

**Recommended Readings**


**References**


Bollen, K. A. (1984). Multiple indicators: Internal con-
Reliability and Construct Validation

491


Gasull, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big Five personality do-
main. *Journal of Research in Personality*, 37, 504-528.


Reis, H. T., & Gable, S. L. (2000). Event-sampling and other methods for studying everyday experience. In H. T. Reis & C. M. Judd (Eds.), Handbook of research methods in social and personality psychology (pp. 190-222). New York: Cambridge University Press.


