WARNING CONCERNING COPYRIGHT RESTRICTIONS

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproduction of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If electronic transmission of reserve material is used for purposes in excess of what constitutes “fair use”, that user may be liable for copyright infringement.


© Cambridge University Press 2000
CHAPTER THIRTEEN

Measurement: Reliability, Construct Validation, and Scale Construction

OLIVER P. JOHN AND VERONICA BENET-MARTÍNEZ

Is an alpha reliability of .70 high enough? How do I know my questionnaire scale is unidimensional? What do I need to do to show that my measure is valid? These are the kinds of questions that methodologists are often asked. The answers to these questions are important for everybody who does empirical research in social-personality psychology, and they all involve basic issues in measurement. Yet, when we teach courses on measurement and test construction, we seldom encounter much enthusiasm. In fact, most students think that measurement is outright boring. However, without measurement there would be no empirical science.

Consider an unusual and extreme but illustrative example: recent research on the vast societal problem of child molestation (see Harris & Rice, 1996). Theory suggested a crucial variable: Child molesters may sexually prefer and be responsive to children, whereas nonmolesters are more responsive to adults. The researchers were faced with what seemed to be insurmountable measurement problems — how could they measure a construct such as “sexual responsiveness to children” in individuals who had reason to deny to others and themselves that they are attracted to children? Self-report did not seem a viable option when studying sex offenders and sexual aggression. Eventually, the researchers developed an ingenious phallicometric procedure that allowed them to measure genital blood flow in response to slides depicting pictures of nude adults and children. Even though they had a seemingly fool-proof physiological measure, the investigators took painstaking care to attend to measurement issues: Do the blood-flow measurements generalize across equivalent kinds of pictures? Do they replicate over time and testing situations? Do they validly differentiate groups of known offenders from (presumed) nonoffenders? Do they converge with measures obtained with other methods, such as reports from clinicians, and would they predict future abuse?

SOME GENERAL CONSIDERATIONS
IN MEASUREMENT

These questions — traditionally discussed under the headings of reliability and validity — 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 to other samples, items, measures, methods, outcomes, and so on (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). If we cannot make such generalizations, our measurements are obviously much less useful than if we can provide explicit evidence for generalizability.

The notion of generalizability also reminds us that good measurement implies not only that we can reproduce or replicate the same measurement but also 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 phallicometric example, the blood-flow measurements

The preparation of this chapter was supported, in part, by Grant MH42555 and 43948 from the National Institute of Mental Health and a sabbatical award from the University of California, Berkeley, to the first author. We are indebted to Harry Reis and Chick Judd for their enormous editorial efforts on behalf of this chapter and to Lewis R. Goldberg and Richard W. Robins for their helpful comments on an earlier draft. Correspondence may be addressed to Oliver P. John, Department of Psychology, University of California, Berkeley, CA 94720-1650 (e-mail: ojohn@socrates.berkeley.edu).
would be useless if they failed to help us understand differences between offenders and nonoffenders. Another basic idea implicit in the idea of generalizability is that all psychological measurement—self-reports, observer ratings, even physiological measures—is 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., there may have been something about the particular slide used, the blood-flow meter may have shifted slightly, etc.), and the resulting observation (or score) is only imperfectly related to what we want to measure, namely sexual responsiveness. To counteract this limitation of single measurements, psychologists obtain multiple measurements (e.g., across different stimuli, experimenters, or observers) and then aggregate them into a more generalizable composite score.

Like most models, measurement models (e.g., tests, scales, or variables) 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 successively 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 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.

Defining Measurement as Building and Evaluating Models

Admonished that psychologists often talk “at great length” about phenomena and concepts they have not defined (Dawes & Smith, 1985, p. 509), we shall briefly consider what measurement is and how it may be defined. An early definition comes from Stevens (1951), one of the founders of measurement theory, who suggested that measurement is the assignment of numbers to objects or events according to rules. However, it is now generally agreed that measurement requires more than that, and Dawes and Smith, Himmelfarb (1993), and Judd and McClelland (1998) present excellent discussions of the relevant historical and conceptual issues. We agree that it is most useful to think of measurement as the process of building models that represent the phenomena of interest, typically in quantitative form. Judd and McClelland (1998) articulated this point of view:

The raw data...of the social and behavioral sciences...consist of infinitely minute observations of ongoing behavior and attributes of individuals, social groups, social environments, and other entities or objects that populate the social world. 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 (pp. 181, emphasis added).

Psychometric and Representational Approaches to Measurement

The present chapter focuses on what has become known as the psychometric or nonrepresentational approach to measurement. Representational measurement has been discussed in several extensive reviews, especially in the context of attitude measurement (Dawes & Smith, 1985; Himmelfarb, 1993). In brief, the basic assumption of representational measurement is that numbers are assigned to entities such that the properties of the numbers (e.g., “greater than,” “multiplication”) represent empirical relations. A good example of representational measurement is the Mohs Scale of Hardness, which measures the hardness of rocks in terms of an ordinal scale (Dawes & Smith, 1985, p. 532): Rock X is harder than Rock Y if and only if X can scratch Y. The key feature of representational measurement in this example is the empirical relation that can be shown to exist between any pair of rocks and that can be represented by the “greater-than” relation among real numbers. One advantage of these kinds of measurement models is that they make predictions about the behavior of the individual entities being measured and thus provide internal consistency checks that can be used to disconfirm the model. For example, the ordinal hardness scale for rocks has to follow the transitivity rule, such that if Rock X is harder than Rock Y and Y is harder than Z, then X has to be harder than Z, and this prediction can be verified empirically by checking whether X does indeed scratch Z.
In contrast, the psychometric approach does not afford such internal consistency checks. For example, although the responses participants make on rating scales are often assigned numbers (e.g., 1 = disagree strongly and 5 = agree strongly), these numbers are not imbued with strong representational meaning that would permit consistency checks (see Dawes & Smith, 1985, for samples and a discussion of rating scales). Instead, the psychometric approach relies on aggregate patterns of data to evaluate a proposed measurement model. It does so because it assumes that each individual response or observation is so prone to error that consistency checks at this level of measurement are simply not meaningful and informative. For example, consider the two self-report items “I am a generous person” and “I am a stingy person” (Hampson, 1998). Although responses to these two items tend to be negatively correlated (i.e., most respondents claim one of the two traits but not both), the correlations are not even close to −1.0, the number we would expect if people were semantically consistent. Instead, only some people are very consistent; there are vast individual differences, even among college students, and more verbally intelligent students show greater consistency (Goldberg & Kilkowski, 1984). In other words, people are not like rocks—they are much less consistent in their behavior, scratching or otherwise. Thus, the psychometric approach tends to ignore consistency checks at the level of the individual and instead relies on patterns of variances and covariances that reflect relations at the aggregate level in probabilistic form (e.g., in this sample, individuals who gave relatively high ratings to “generous” were unlikely to give high ratings to “stingy”).

Although representational measurement promised to provide a strong and defensible foundation for psychological measurement, it has so far failed to deliver on that promise. During the 1970s and 1980s a slew of studies, inspired by Tversky and Kahneman’s (1974) pioneering work, showed that people’s preferences, risk perceptions, political attitudes, and so on often violate the transitivity rule required for ordinal scaling and that judgments may shift substantially depending on the framing of the questions or items. Dawes and Smith (1985) noted that “representational measurement is rare in the field of attitude; instead, this field is permeated by questionnaires and rating scales” (pp. 511–512). More recently, Cliff (1992) called representational measurement “the revolution that never happened” (p. 186), and Dawes (1994) concurred. For these reasons, and because Judd and McClelland (1998) provided an excellent up-to-date review and discussion of the representational approach, the present chapter is devoted to the psychometric approach.

Overview

The remainder of this chapter is organized into three parts. We begin with the historically early conceptions of reliability and then move to more recent and increasingly complex views that emphasize construct validation and model-testing as a broader, more integrative approach. In the first part of this chapter, we consider issues traditionally discussed under the heading of reliability; review several still persistent definitions or “types” of reliability coefficients; discuss in some detail the problems and misuses of coefficient alpha, the most commonly used psychometric index in social-personality psychology; and then suggest generalizability theory as a broader and more heuristic perspective. In the second part, we examine issues related to construct validation, beginning with early definitions and designs to establish validity, followed by a broader view that considers construct validation as the crucial issue in psychological measurement and includes a broad range of validity evidence, focusing on convergent and discriminant aspects. In the third part, we consider model-testing in construct validation and scale construction. After a brief introduction to measurement models in structural equation modeling (SEM), we discuss an empirical example that reexamines the issue of dimensionality as an aspect of structural validity, and then we consider issues in questionnaire construction, reviewing three classical strategies (external–criterion, rational–intuitive, and internal–factor analytic) and suggesting an integrated model adopting the construct-orientated approach.

It sometimes seems that methodologists write papers that are of great interest to other methodologists. Instead, the present chapter is focused on what in our experience has proven useful and of interest to graduate and postdoctoral students, with the goal of devising sound measurement models and evaluating them, rather than covering mathematical formulae or statistical derivations. Respecting the glacial pace of methodological advances (Cohen, 1990), we do discuss current practice, even when it is outdated, and then point to more recent conceptualizations. Finally, whenever possible, we have avoided technical language, omitted Greek symbols, and used examples to make this chapter as concrete and accessible as possible.
RELIABILITY AND GENERALIZABILITY

It should by now be obvious that most measurement procedures in psychology are subject to "error." Many different sources may contribute to such error. In the social-personality literature, 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 rater to obtain their judgments; the instructions, definitions, and questions on these forms may be difficult to understand or require complex discriminations, again entering error into the measurement. In short, characteristics of the participant, the testing situation, the test or instrument, and the experimenter can all affect reliability.

Reliability refers to the consistency of a measurement procedure, and indices of reliability describe the extent to which the scores produced by the measurement procedure are reproducible. Consider the example of a bathroom scale; if it gives different readings in three successive weighings of the same person, we would hardly call the scale reliable.

Classical Test Theory

Issues of reliability have traditionally been treated within the framework of classical test theory (Guilford, 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 $T$ (e.g., the person's actual weight). This seemingly simple formulation, that each measurement can be partitioned into a true score, $T$, and measurement error, $e$, is the fundamental assumption of classic test theory. Conceptually, each true score represents the mean of a very large number of measurements on 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 will return), and it is this assumption that permits the definition of error in statistical terms.

All conceptions of reliability involve the notion of repeated measurements. Classical test theory has relied heavily on the notion of parallel tests—that is, tests that have the same mean, variance, and distributional characteristics, and that correlate equally well with other 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:

$$
\text{Variance} (X) = \text{Variance} (T + e) \\
= \text{Variance} (T) + \text{Variance} (e)
$$

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:

$$
\text{Reliability} = \frac{\text{Variance} (T)}{\text{Variance} (X)} \\
= 1 - \frac{\text{Variance} (e)}{\text{Variance} (X)}
$$

In other words, if there is no error, reliability would be $1$; if there is only error and no true-score variance, reliability would be $0$. The correlation between the true variable and the observed score is the square root of the reliability.

Specific Types of Reliability Evidence

Because classical test theory defined parallel test in purely mathematical terms, it provided little substantive specification or restriction of the types of measurement procedures that might be considered parallel. Beginning in the 1950s, several experimental designs were distinguished, and they are summarized in Table 13.1: Retest (or stability), equivalence, and internal consistency (or split-half). These distinctions are meant to convey the idea that "reliability is a general term referring to many types of evidence" (American Psychological Association [APA], 1954, p. 28). Each of the different designs spelled out in Table 13.1 takes account quite different sources of error. Retest (or stability) designs estimate how much responses vary with individuals across time and situation, thus reflect error due to differences in the situation and conditions of test administration or observation.1

1 As noted in Table 13.1, both Pearson and interna-
TABLE 13.1. Reliability: Facets of Generalizability, Traditional Definitions of Reliability Coefficients, and Estimation Procedures

<table>
<thead>
<tr>
<th>Facet of generalizability</th>
<th>Major sources of error</th>
<th>Traditional reliability coefficient</th>
<th>Procedure</th>
<th>Statistical analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Times</td>
<td>Change of participant’s responses over time; change in testing situation</td>
<td>Retest (or stability)</td>
<td>Test participants at different times with same form</td>
<td>Pearson or intraclass correlation</td>
</tr>
<tr>
<td>Forms</td>
<td>Differences in content sampling across “parallel” forms</td>
<td>Equivalence</td>
<td>Test participants at one time with two forms covering same content</td>
<td>Pearson or intraclass correlation</td>
</tr>
<tr>
<td>Items</td>
<td>Content heterogeneity and low content saturation in the items</td>
<td>(a) Split-half (b) Internal consistency</td>
<td>Test participants with multiple items at one time</td>
<td>(a) Correlation between test halves (Spearman-Brown corrected) (b) Coefficient alpha</td>
</tr>
<tr>
<td>Judges or observers</td>
<td>Disagreement among judges</td>
<td>Internal consistency</td>
<td>Obtain ratings from multiple judges on one form and occasion</td>
<td>(a) Pairwise interjudge correlation (b) Coefficient alpha (c) Intraclass correlation</td>
</tr>
</tbody>
</table>

Procedures estimate error due to different content-sampling and item-selection in two alternate forms of the test. *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.

**Coefficient Alpha: Ubiquitous but Not a Panacea**

We now consider coefficient alpha (Cronbach, 1951) because this internal consistency index plays such an important role in the social–personality literature. A perusal of the articles published in the *Journal of Personality and Social Psychology* and *Personality and Social Psychology Bulletin* shows that de facto alpha is the index of choice when authors want to claim that their measure is reliable. Moreover, contrary even to the recommendations in the *Standards* (APA, 1985), alpha is usually the only reliability evidence considered.

Why has alpha become the Bill Gates of measurement reliability? We suspect it is the relative ease with which alpha is both obtained and computed. Alpha does not require collecting data at two different times from the same participants, as retest reliability would, or the construction of two alternate forms of a scale, as parallel-form reliability would require. Alpha is the “least effort” reliability index: it can be used as long as the same participants responded to multiple items thought to indicate the same construct. And computationally, today’s SPSS and statistical software packages allow the user to view the alpha of many alternative scales formed from any motley collection of items with just a few mouse clicks. However, although alpha has many important uses, it also has some important limitations. Although long known to methodologists, these limitations are often underappreciated by researchers and are therefore reviewed here in some detail.

**TWO DETERMINANTS OF ALPHA.** Cronbach’s (1951) alpha is a generalization of split-half reliability, representing the mean of the reliabilities computed from all possible split halves of the test. As such, alpha is a function of two parameters: (a) the interrelatedness of the items in a test or scale and (b) the length of the test. Consider Table 13.2, which shows the interitem correlation matrices for two hypothetical tests,
TABLE 13.2. Interitem Correlation Matrices for Two Hypothetical Tests with the Same Coefficient Alpha Reliability of .81

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.3</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.3</td>
<td>.3</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>—</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.6</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.6</td>
<td>.6</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>—</td>
</tr>
</tbody>
</table>

The idea that test length can compensate for lower levels of interitem correlation is formalized in the Spearman–Brown prophecy formula, which specifies the relation between test length and reliability (see, e.g., Lord & Novick, 1968). Given a particular level of mean interitem correlation, the Spearman–Brown formula allows the researcher to derive the number of items needed to achieve a certain level of alpha. Figure 13.1 shows this relation for mean interitem correlations of .20, .40, .60, and .80 in graphic form. Three points are worth noting. First, the alpha reliability of the total scale always increases as the number of items increases (as long as adding items does not lower the mean interitem correlation). Second, the utility of adding ever more items diminishes quickly, so that adding the 15th item leads to a much lesser increase in alpha than adding the 5th item, just like consuming the 15th chocolate bar or beer adds less enjoyment than did the earlier ones. Third, less is to be gained from adding more items if those items are highly intercorrelated (e.g., .60) than when they show low content saturation (e.g., .20). The lesson here is that we need to be careful in interpreting alpha: We cannot interpret empirical findings without considering scale length. In contexts where the researcher is interested in the homogeneity of the items, a direct index of item content saturation (e.g., the mean interitem correlation) may be more informative than alpha.

**ALPHA DOES NOT INDEX UNIDIMENSIONALITY.** The examples in Table 13.2 also illustrate a second issue with alpha: Contrary to popular belief, alpha does not measure the homogeneity of the interitem intercorrelations, nor does it indicate that a scale is unidimensional. In fact, although Tests A and B in Table 13.2 have the same alpha, they differ radically in the homogeneity (vs. dispersion) of the correlations among their items. For Test A, they are completely homogeneous (all are .3, with a standard deviation (SD) of 0 in this hypothetical example), whereas for Test B they vary considerably from .3 to .6, with an SD of .15. Because alpha does not represent this variability, Cortina (1993) derived an index that reflects the spread of interitem correlations and argued that this index should be reported along with alpha. A large spread in interitem correlations is a bad sign because it suggests that either the test is multidimensional or the interitem correlations are distorted by substantial sampling error. In the example, the pattern of item intercorrelations for Test B suggests that the problem here is multidimensionality. Clearly, the responses to these six items are a function of not one, but two, factors: Items 1, 2, and 3 correlate much more substantially (mean r = .6) with each other than they correlate (mean r = .3) with items

2 Consistent with Schmitt (1996), the examples are presented in correlational terms (rather than in covariance terms) simply for ease of interpretation and convenience. Alphas are in fact standardized alphas (i.e., after standard scoring all variables).

3 Cortina’s (1993) index should not be confused with the standard error of alpha, which can be computed under certain distributional assumptions (cf. Feldt, Woodruff, & Salk, 1987).
of .42, which is the mean interitem correlation) were higher than for the truly unidimensional Test A. As expected, a two-factor model significantly increased fit for Test B, and perfect fit was obtained when we specified a model with two correlated factors. Reflecting their .60 correlation with each other, items 1, 2, and 3 loaded .775 on factor 1 and 0 on factor 2, whereas items 4, 5, and 6 loaded 0 on factor 1 and .775 on factor 2; the mean intercorrelation of .3 between the items in these two sets gave rise to a .50 correlation estimated between the two latent factors.

It is important to emphasize that the issue of error (or unreliability) present in an item is separate from the issue of multidimensionality (which is discussed in later sections on structural validity). In the SEM analyses summarized above, the item loadings represent how much of the item variance is shared across items (thus generalizable), whereas 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, on the other hand, is captured by the relative fit of the one-factor model over multiple-factor models. Thus, comparing again Tests A and B in Table 13.2, the longer Test A is clearly more unidimensional than Test B, 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. We return to these issues later in this chapter when we discuss the measurement model in SEM.

Once we know that a test is multidimensional, can we go ahead and still use alpha as a reliability index? Unfortunately, the answer is no. As Cronbach (1947, 1951) recognized early on, we can estimate the reliability of a multidimensional test or scale only through parallel forms, and the two parallel forms must show the same factor structure. 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 multidimensional, one should score two
unidimensional subscales and then use alpha to index their reliabilities separately.4

HOW LARGE SHOULD ALPHA BE? IT DEPENDS ON THE CONSTRUCT. Students often ask questions like “my scale has an alpha of .70—is that good enough?” and they are frustrated when the answer is “that depends.” Although it would be nice to have a simple cookbook for measurement decisions, there is no particular level of alpha that is necessary, adequate, or even desirable in all contexts. Although Nunnally (1978) suggested that “reliabilities of .7 or higher will suffice” (p. 245), an alpha of .70 is not a benchmark every scale must pass. It is easy to find examples in the literature that use this arbitrary standard. For example, Gray-Little, Williams, and Hancock (1997) noted that alphas between .72 and .88 are usually taken to indicate “acceptable to high reliability” (p. 444). However, as we have seen above, alpha needs to be interpreted in terms of its two main parameters — interitem correlation as well as scale length — and in the context of how these two parameters fit the nature and definition of the construct to be measured. In any one context, a particular alpha may be just right, or too low, or too high. As Pedhazur and Schmelkin (1991) put it,

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)

The definition of the construct to be measured is a crucial parameter in interpreting alpha. Consider a researcher who wants to measure the broad construct of extraversion, which includes sociability, assertiveness, and talkativeness, and has constructed a scale with the following items: “I like to go to parties,” “Parties are a lot of fun for me,” “I do not enjoy parties,” (reverse scored) and “I’d rather go to a party than spend the evening alone.” Note that these items are essentially paraphrases of each other and represent the same item content (liking parties) stated in slightly different ways. Cattell (1972) called these kinds of scales “bloated specifics” — they have high alphas simply because the item content is so redundant and interitem correlations are very high. Thus, alphas in the high .80s or even .90s, especially for short scales, may not indicate an impressively reliable scale but instead signal redundacy or narrowness in item content.

For example, the 10-item Rosenberg (1979) Self-Esteem Scale has alphas approaching .90 in student samples, and the pairwise correlations between some items approach .70 (Gray-Little et al., 1997). Some of these self-esteem 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 Robins & Hendin, 1999).

This phenomenon is also known as the attenuation paradox because increasing the internal consistency of a test beyond a certain point will not enhance construct validity and may even come at the expense of validity when the added items emphasize one part of the construct (e.g., party-going) over other important parts (e.g., assertiveness). This paradox emphasizes an important point in this chapter: Our goal in measurement is to maximize validity rather than internal consistency, and issues of meaning and conceptualization play a key role in all decisions about measurement.

The party-going items on our imaginary extraversion scale illustrate how easy it is to boost alpha by adding redundant items to a scale. However, unless one is specifically interested in party-going behavior, 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 sociability and even less useful as a measure of the broader construct of extraversion. Although the scale may predict the frequency of party attendance with great precision (or fidelity), it is less likely to relate to anything else of interest because of its narrow bandwidth. Conversely, broad-bandwidth measures (e.g., an extraversion scale or a general attitude measure of conservatism) can predict a wider range of outcomes or behaviors but do so with lower fidelity. This phenomenon is known as the bandwidth-fidelity trade-off (Cronbach & Gleser, 1957) and has proven to be of considerable importance in the literature on attitudes (Eagly & Chaiken, 1993; Fishbein & Ajzen, 1974) and personality traits (Epstein, 1980; John, Hampson, & Goldberg, 1991). In general, then, the attitude or trait serving as the

---

4 There is an important exception where this internal-consistency conception does not apply. In most social-personality measurement, the indicators of a construct are seen as effects caused by the construct; individuals endorse particular attitude statements because of underlying individual differences in attitudes. 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.
predictor should be measured at a similar level of abstraction as the criterion to be predicted, so that predictive relations are not going to be attenuated.

The close connection between the hierarchical level of the construct to be measured and the content homogeneity of the items is illustrated in Figure 13.2. Sociability, assertiveness, and talkativeness are three scales that are positively intercorrelated and together define the broader construct of extraversion. (Our earlier example of the party-going scale might be represented as an even lower-level scale, representing one of the components of the sociability scale.) Consider the assertiveness scale. Because its six items are selected to represent a narrow range of content (e.g., all the assertiveness items have to do with dominance and self-assertion in various situations), item content should be relatively homogeneous, leading to a substantial mean interim correlation. Now consider an equally long extraversion scale (shown on the right side of Figure 13.2), made up of two sociability items, two assertiveness items, and two talkativeness items. Compared with the lower-level scales, the item content on this scale is much more heterogeneous, leading to a lower mean interim correlation and thus a lower alpha.

One implication of Figure 13.2 is that if one wants to measure broader constructs, one should probably include a larger number of items to compensate for the greater content heterogeneity; for example, one might use all 18 items to measure the superordinate extraversion construct defined on the left side of Figure 13.2.

**Figure 13.2.** Relation between hierarchical level and content homogeneity in interpreting the size of alpha reliability.

This principle underlies the construction of many hierarchically organized assessment instruments, such as in the literature on the self-concept (e.g., Marsh, Byrne, & Shavelson, 1992) and on personality traits. For example, in Costa and McCrae’s (1992) NEO PI-R, each of the Big Five personality dimensions is defined by six more specific “facet” scales, which, in turn, are each measured with eight items; the resulting 48-item superordinate Big Five scales all have reliabilities exceeding .90.

**Correcting for Attenuation**

According to classical test theory (e.g., Lord & Novick, 1968), researchers should be concerned about reliability 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. For example, for a reliability of .70, the expected upper limit for a correlation would be the square root of .70, namely .84, and for a reliability as low as
.60, the limit would be .77. These numerical examples show that lower reliabilities can reduce estimates of external correlations and, everything else equal, higher reliabilities are to be preferred.

However, other properties of the instrument need to be considered as well in planning one’s research. As Burisch (1984, 1986) has shown, “short scales not only save testing time but also avoid subject boredom and fatigue . . . there are subjects . . . from whom you won’t get any response if the test looks too long.” Consider, for example, the 48-item long Big Five trait scales from the NEO PI–R (Costa & McCrae, 1992). Despite their reliabilities above .90, they are used less frequently by other researchers than are shorter scales that have reliabilities in the .70s and .80s. When participants time and attention is at a premium, the trade-off between length and reliability may well be worth it – note that the drop in reliabilities from .90 to .80 and to .70 lowers the upper limits for external correlations only from .95 to .89 and to .84 for the shorter scales.

Researchers sometimes use reliability indices (typically alpha) to correct observed correlations between two measures for attenuation due to unreliability. Such corrections are sometimes used to estimate the correlation between the latent constructs underlying their measures (see also the section on SEM below) – that is, what would the correlation be if both measures were assessed with perfect reliability? This can be useful to compare effect sizes across variables or studies. Another application is in contexts where researchers want to distinguish the long-term stability of attitudes or personality from the reliability of measurement or compare stability estimates for different groups, such as men and women (J. Block & Robins, 1993). The correction formula (Cohen & Cohen, 1975; Lord & Novick, 1968) is simple: Divide the observed correlation by the square root of the product of the two reliabilities. This correction expresses the size of the association relative to the maximum correlation attainable given the imperfect reliabilities of the two measures.

However, the ease of this correction should not lead to sloppy measurement. Appealing to the relative brevity of one’s measures to excuse low reliability is, as we have seen above, not the only explanation for low alphas, and certainly not an excuse. In many situations, low reliability will create problems for estimating effect sizes and testing hypotheses. 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.

### Reporting Basic Psychometric Data

Whereas most social–personality researchers do report alpha reliabilities for their scales, few report their intercorrelations. This information is often crucial, for example, when multiple scales are scored from the same data source (e.g., a self-report attitude measure) and when the research question implies relative independence among the constructs measured. For example, intercorrelations among predictors are important for understanding the results of multiple regression analyses (e.g., see the numerical examples provided by Goldberg, 1991), and concerns have been raised about the correlations among constructs postulated to be conceptually uncorrelated, such as the Big Five personality dimensions (e.g., J. Block, 1995; John & Srivastava, 1999). Thus, we agree with Schmitt (1996), who argued that, at the very least, research reports should regularly present a matrix that includes reliability information for the key measures (on the diagonal), the intercorrelations among these measures (below the diagonal), and probably also the intercorrelations corrected for attenuation due to unreliability (above the diagonal). Table 13.3 gives an example from our own research on the

### Table 13.3. How to Report Simple Psychometric Information for Multiple Scales: Alpha Coefficients, Observed Correlations, and Corrected Correlations among the Big Five Inventory (BFI) Scales

<table>
<thead>
<tr>
<th>Scales</th>
<th>E</th>
<th>A</th>
<th>C</th>
<th>N</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraversion (E)</td>
<td>.88</td>
<td>.17</td>
<td>.28</td>
<td>−.34</td>
<td>.30</td>
</tr>
<tr>
<td>Agreeableness (A)</td>
<td>.14</td>
<td>(.79)</td>
<td>.34</td>
<td>−.38</td>
<td>.06</td>
</tr>
<tr>
<td>Conscientiousness (C)</td>
<td>.24</td>
<td>.27</td>
<td>.82</td>
<td>−.22</td>
<td>.10</td>
</tr>
<tr>
<td>Neuroticism (N)</td>
<td>−.29</td>
<td>−.31</td>
<td>−.18</td>
<td>(.84)</td>
<td>−.17</td>
</tr>
<tr>
<td>Openness (O)</td>
<td>.25</td>
<td>.05</td>
<td>.08</td>
<td>−.14</td>
<td>(.81)</td>
</tr>
</tbody>
</table>

Note: $N = 711$ U.S. college students. Data from Benet-Martínez and John (1998). Alpha coefficients are presented on the diagonal, observed correlations below the diagonal, and correlations corrected for attenuation above the diagonal.
Big Five Inventory scales (John, Donahue, & Kentle, 1991; John & Srivastava, 1999). The inclusion of the uncorrected intercorrelations allows the reader to evaluate the size of the reliability coefficients relative to the overlap among the scales; reliabilities should be substantially larger than these intercorrelations. The inclusion of corrected intercorrelations is helpful because it removes differences among the intercorrelations that are simply due to differential reliability, thus making comparisons among intercorrelations much easier. The corrected coefficients are also useful for identifying pairs of scales that lack discrimination— that is, they are so highly intercorrelated that postulating two separate underlying constructs is not sensible either theoretically or practically.

Beyond Classical Test Theory:
Generalizability Theory

The distinctions among “types of reliability” emphasized in the literature and summarized in Table 13.1 had a number of unfortunate consequences. First, they 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; reliability depends on the particular facet of generalization being examined (Cronbach, Rajaratnam, & Gleser, 1963). Second, 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. Third, 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 (Shavelson, Webb, & Rowley, 1989).

Therefore, the APA (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 sufficiently over the years, and generalizability theory has not fully replaced these more simplistic notions. To emphasize that the classical conception of random error is outdated, the very first column in our Table 13.1 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, as shown by the last row in Table 13.1, concern with interjudge agreement may actually be a concern with the question of how accurately we can generalize from a given set of ratings to ratings by another set of judges. Or we might 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 might want to test the generalizability of a scale originally developed in English to a Chinese language and cultural context.

All these facets of generalizability represent legitimate research concerns that we will reconsider later in this chapter under the heading of construct validation; 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 13.1, we should use more general estimates, such as intraclass correlation coefficients (see Shrout & Fleiss, 1979), to probe particular aspects of the dependability of measures. For example, the intraclass correlation coefficient (see Judd & McClelland, 1998, for numerical examples) can be used to index the generalizability of one set of judges to a universe of similar judges.

It is perplexing: Generalizability theory should hold considerable appeal for personality–social 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., Shavelson et al., 1989), generalizability theory is not used as widely as it should be. A recent exception is the flourishing research on determinants of consensus among personality raters (Kenny, 1994; see also Kashy & Kenny, this volume, Ch. 17) and the determinants of self–other agreement (John & Robins, 1993).

Generalizability theory is especially useful when data are collected in nested designs and multiple facets may influence reliability. A nice illustration is King
Figueroedo’s (1997) study of chimpanzee personality differences. They 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; fortunately for their purposes, setting and subject variables turned out to be unimportant. It is a shame 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.

**Item Response Theory**

The measurement model and procedures of classical test theory have also been criticized by psychometricians advocating item response theory (IRT) as an alternative and more advanced approach (Embretson, 1996; Mellenbergh, 1996). In the classical conception of reliability, the characteristics of the individual test-taker and the characteristics of the test cannot be separated (Hambleton, Swaminathan, & Rogers, 1991). That is, the person’s standing (or level) 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. Furthermore, the psychometric characteristics of the test depend on the particular sample of respondents being measured: 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 an undergraduate student sample but fail to make reliable distinctions among bible-belt evangelists. In short, classical test theory is not helpful 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 error of measurement 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.

These limitations can be addressed in IRT, which describes the relation between individuals’ responses to a particular item and the construct underlying those responses with a function called the *item characteristic curve*. This curve depicts the probability that individuals at different levels of the construct would endorse the item; it thus provides information about how well the item discriminates those with high versus low levels of the underlying trait and also about how difficult the item is. This information is particularly useful to researchers interested in detecting biases in their items; according to IRT, an item is an unbiased measure of a construct, say conservatism, if individuals who are equally conservative have the same expected score on the item, regardless of conceptually unrelated memberships in gender, ethnic, or cultural groups.

In the context of constructing and evaluating scales and other multi-item measures, IRT procedures have two attractive features. First, they permit researchers to select items on the basis of both difficulty and discrimination, rather than relying on the item-total correlations offered by classical test theory. Second, IRT procedures can be used to assess a person’s standing on the construct without having to administer the entire scale, a procedure known as computerized adaptive testing (Waller & Reise, 1989).

Until recently, IRT was limited computationally to dichotomous (true–false) response formats and unidimensional constructs. It was therefore much more useful for educational and achievement research (where item difficulty has an inherent psychological meaning) than for social-personality research, which relies heavily on multistep rating scales. However, with the recent extensions of IRT and IRT software to rating scales and multidimensional models (Kelderman & Rijkse, 1994), IRT’s “new rules of measurement” (Embretson, 1996, p. 341) may soon appear more frequently in our journals. For example, Gray-Little et al. (1997) used IRT to explore the properties of the 10 items on the Rosenberg Self-Esteem Scale. Results indicated that the 10 items indeed define a unidimensional trait. However, given the uniformity of the item discrimination parameters, the scale could easily be shortened without compromising the measurement of global self-esteem, a conclusion consistent with our earlier discussion of construct definitions and item redundancy (Robins & Hendin, 1999). The IRT analyses also indicated that the items discriminate better at low and moderate levels of self-esteem than at higher levels.
With its current selection of items, the scale may fail to
differentiate reliably between truly high levels of self-
esteem and narcissistically exaggerated, grandiose self-
views (John & Robins, 1994).

More generally, then, IRT provides quantitative pro-
cedures to describe the relation of a particular item to
the latent construct being measured in terms of diffi-
culty and discrimination parameters. This information
can be useful for item analysis and scale construction,
permitting researchers to select items that best mea-
sure a particular level of the construct of interest and
detecting items that are biased for particular groups of
individuals.

To summarize, in this section we focused on classi-
cal test theory approaches to reliability, specific types
of reliability indices, issues with coefficient alpha (test
length, unidimensionality, and construct definitions),
and the practice of correcting for attenuation. In dis-
cussing these issues, 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 raise complex conceptual issues and high-
light that the meaning and interpretation of measure-
ments is crucial to evaluating the quality of our mea-
surements. Traditionally, issues of score meaning and
interpretation are discussed under the heading of va-
lidity. We focus on the validity of measured variables
here; the validity of manipulated variables is discussed
in Chapters 1 and 2 of this volume.

CONSTRUCT VALIDATION

Traditional Definitions of Validity

As described by Cronbach and Meehl (1955), the
APA committee on psychological tests initially distin-
guished among several types of validity, which are
given in the top part of Table 13.4. Content validity
is established by demonstrating that the items are a
representative sample of the universe of item content
relevant to the construct. This aspect of validity is typi-
cally established deductively; first the investigator de-
"fines a universe of items (i.e., a hypothetical set of all
possible kinds of relevant item content) and then sam-
...
TABLE 13.4. Types of Validity and Validity Evidence: Major Approaches

Early Approaches (e.g., Cronbach & Mehl, 1955)

**Content validity:** Extent to which the items are a representative sample of the behavior domain to be measured

**Face validity:** Extent to which the items appear to measure the intended construct

**Criterion-oriented (or external) validity:**
- (a) Predictive: Extent to which an individual’s future score on a criterion is predicted from prior test scores
- (b) Concurrent: Extent to which the test scores estimate an individual’s present criterion score

**Construct validity:** Whether the measure accurately reflects the construct intended to measure

- Elaboration of Construct Validity (e.g., Loevinger, 1957; Messick, 1989)

**Content validity:** Evidence of content relevance, representativeness, and technical quality of items

**Substantive validity:** Evidence for response consistencies or performance regularities that are reflective of domain processes

**Structural validity:** Evidence for internal structure of the scores that is consistent with the internal structure of the construct domain

**Generalizability:** Evidence for score properties and interpretations that generalize to and across population groups, settings, and tasks

**Consequential validity:** Rationale and evidence for evaluating the intended and unintended consequences of score interpretation and use, including test bias and fairness

**External validity:** Convergent and discriminant evidence from multitrait multimethod comparisons, as well as criterion relevance

Examples of Validation Procedures

**Expert judgments and review:** Test whether experts agree that items are relevant and represent construct domain; use ratings to assess item characteristics, such as comprehensibility and unambiguity

**Differentiation between criterion (or contrast) groups:** Test size and direction of expected differences between groups on the construct of interest

**Factor analysis:** Test hypothesized structure of the construct domain (e.g., whether items thought to define the construct load on the same factor and not on other factors)

**Correlation:** Test relation between measure of construct and measure of other distinct constructs

**Multitrait multimethod:** Test whether different measures of the same construct correlate more highly than measures of different constructs using same and different methods (e.g., instruments, data sources, languages)

---

collect the data). In this view (e.g., Judd & McClelland, 1998), scores on observed variables potentially reflect three sources of variance: (a) the construct we intend to measure (convergent aspects of validation), (b) a variety of other constructs (or sources of influence) we would like to avoid measuring (discriminant aspects of validation), and (c) random error (or unreliability). This broad construct view thus highlights convergent and discriminant validity and considers reliability as just another piece of evidence for the construct validity of the proposed measurement.

Messick (1989, 1995) has articulated a comprehensive program of construct validation that addresses the meaning of test scores in test interpretation and use. His view highlights that validity is an “integrative evaluative judgment of the degree to which evidence and theoretical rationales support the adequacy and appropriateness” of the theoretical specification of the construct (Messick, 1989, p. 13). Thus, validity is considered a property of the interpretation of a measure, rather than a property of the measure itself; for example, there may be substantial evidence to support the interpretation of a particular attitude scale as a measure of individual differences in liberal values but no validity evidence for its interpretation as a measure of intelligence or extraversion. Of course, if the theoretical account of the construct is specified clearly and in detail, specific predictions about relations to other constructs and criteria can be readily made, thus simplifying the process of collecting evidence that supports or disconfirms a particular interpretation of the test score.
Like any other theory or model, the validity of the particular score interpretation can never be established but is always evolving to form an ever-growing “nomological network” of validity-supporting relations (Wiggins, 1973). Given that multiple pieces of evidence will accumulate to support the hypothesized construct, it is often difficult to summarize the available validity evidence with a simple quantitative index, and investigators have had to resort to qualitative and tabular summaries. For example, Snyder (1987) wrote a whole book to summarize what has been learned about the self-monitoring construct in more than 20 years of empirical research and construct development. More recently, meta-analytic techniques (see Johnson & Eagly, this volume, Ch. 19) have proven useful to make such data summaries more manageable and objective (Schmidt, Hunter, Pearlman, & Hirsch, 1985).

Types of Evidence for Construct Validity

In his integrative account, Messick (1989) specified six forms of evidence that should be sought to examine construct validity. The six forms of evidence are listed and defined briefly in the middle section of Table 13.4.

The first is evidence for 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 are 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. For example, when revising his self-monitoring scale, Snyder (1987) excluded a number of items measuring other-directed self-presentation, thus representing behavioral variability and attitude–behavior inconsistency 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, & Kohnen, 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 positive and negative emotion expression, these scales should not assess variance that must be attributed to theoretically unrelated constructs, such as social desirability or self-esteem (e.g., Gross & John, 1997).

As shown in the third section of Table 13.4, there are a number of validation procedures researchers might use (see also Smith & McCarthy, 1995). Researchers might ask expert judges to review the match between item representation and construct domain specification, and to add or delete items. Another procedure would be to use factor analysis to verify the hypothesized structure of the content domain.

Substantive validity evidence makes use of substantive theories and process models to further support the interpretation of the test scores. Relevant procedures might involve differentiation between criterion (or contrast) groups assumed to differ in the relevant processes. For example, 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 to not need cognition). Even stronger 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 the 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).

Structural validity evidence requires that the correlational (or factor) structure of 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, and we will return to this important issue below in the context of evaluating measurement with structural equations models.
Generalizability evidence, as used here by Messick (1989, 1995), is needed to demonstrate that score interpretations apply across tasks or contexts, times or occasions, and observers or raters. The inclusion of generalizability evidence here makes explicit that construct validity includes consideration of "error associated with the sampling of tasks, occasions, and scorers (that) underlie traditional reliability concerns" (Messick, 1995, p. 746). In this context, we should note that the notion of generalizability encompasses traditional conceptions of both reliability and criterion validity; 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.

As suggested by Figure 13.3, generalizing from a test score to another test constructed according to parallel procedures does increase our confidence in the test but does so only modestly. 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 nontest criteria related to the construct the test was intended to measure. Figure 13.3 thus resembles the layers of an onion, showing how far the test allows us to generalize, with the inner layers representing relatively modest levels of generalization and the outer layers representing further-reaching generalizations to contexts that are more and more removed from the central core (i.e., dissimilar from the initial moment operation).

The kind of validity evidence Messick (1989) considered under the generalizability rubric is crucial in establishing the limits or boundaries beyond which interpretation of the measure cannot be extended. An issue of particular importance for social-personal searchers is the degree to which findings get from "convenience" samples, such as American college students, to groups that are less educated or come from different ethnic or cultural backgounds.

Consequential validity evidence focuses on personal and societal consequences (both intended and unintended) of score interpretation and use. It refers to the test-user to confront issues of test bias and is of paramount importance in contexts where tests are used to make important decisions about individuals. Thus, it is more about validity of use than about the validation of the test per se. Consequential validity is a greater concern in educational and employment settings than in social-personal search contexts, where scale scores and performance in experimental tasks have little, if any, consequence for the research participant. Finally, external evidence covers such a broad range of both convergent and discriminant evidence that we consider it in detail.

**Figure 13.3.** How far can we generalize from a test score? The onion model of generalizability.

**External Validation: Convergent and Discriminant Aspects**

External validity evidence refers to the ability of a test to predict conceptually related behavior, comes, or criteria, and has been emphasized by a wide range of attitude measurement and Smith (1985, p. 512) that "the basis of all measure empirical prediction," his review of personaliters. Wiggins (1973), argued that prediction "is qua non of personality ment." Obviously, it makes sense that a test or scale should construct-relevant criteria, less apparent that we also show that the test does not conceptually unrelated criteria in other words, a full demo of external aspects of...
validation requires requires a demonstration of both what the test measures and what it does not measure.

**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 participant’s need for cognition and with frequency of effortful thinking measured by beepers the participant 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 (1959) 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 example? Certainly, we would expect sizable convergent validity correlations among 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 intercorrelations among the constructs should emerge, regardless of the method used; in other words, the relations among the constructs should generalize across methods.

**METHOD VARIANCE.** One important recognition inherent in the MTMM is that we can never measure the 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). For example, it has been argued that positivity bias in self-perceptions is 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 the positive intercorrelation between these measures may not represent a valid hypothesis about the 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 strengthened considerably if psychological health were measured with a method other than self-report, such as ratings by clinically trained observers (Robins & John, 1997).

**MULTIPLE SOURCES OF DATA: LOTS.** 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 (J. H. Block & Block, 1980).

L data refer to life-event data that can be obtained fairly objectively from the 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 L data are counts of bottles and cans in garbage
containers to measure alcohol consumption (Webb, Campbell, Schwartz, Sechrest, & Grove, 1981) and police records of arrests and convictions to measure juvenile delinquency (Moffitt, 1993).

O data refer to observational data, ranging from observations of very specific aspects of behavior to more global ratings (see Bakeman, this volume, Ch. 6; Kerr, Aronoff, & Messe, this volume, Ch. 7). Examples are careful and systematic observations recorded by human judges, such as in laboratory settings or carefully defined situations; behavior coded or rated from videotape; and 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. A 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. Reaction times are frequently used in studies of social cognition, providing an objective measure of an aspect of performance. An intelligence test is another kind of example. A third is the length of time an individual persists on a puzzle or delays gratification in a standardized situation (Mischel, 1990).

Last, but not least, 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, this volume, Ch. 11), in narratives and life stories (see C. Smith, this volume), and in survey research (Visser, Krosnick, & Lavrikas, this volume, Ch. 9). Daily experience sampling procedures (see Reis & Gable, this volume, Ch. 8) 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 the individual 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 Visser, Krosnick, & Lavrikas, this volume, Ch. 9) and response extremeness (Hamilton, 1968).

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 & John, 1998). 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.

Fortunately, although social–personality psychologists use self-reports most frequently, other methods are available and used. 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, it seems that multimethod designs have been underused in construct validation efforts. In a way, 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 took off in the 1970s when SEM became available and provided powerful analytical tools to estimate separate trait and method factors (Kenny, 1976; Kenny & Kashy, 1992; Schwarzer, 1986; see also Wegener & Fabrigar, this volume, Ch. 16). A number of excellent reviews and overviews have appeared recently. For example, West and Finch (1997, pp. 155–159) provided hypothetical data to illustrate three scenarios: (a) convergent and discriminant validity with minimal method effects, (b) strong method effects, and (c) effects of unreliability and lack of discriminant validity. Judd and McClelland (1998, Tables 13.11–13.15) provide a series of examples that illustrate Campbell and Piske’s (1939) original principles of convergent and discriminant validation as well as the application of SEM techniques to estimate
separate trait and method effects (see Kenny & Kashy, 1992, for specific issues in fitting SEM models).

To summarize, in this section we reviewed Messick’s (1989) six forms of evidence relevant to construct validation (see Table 13.4), and then considered one of them, external validation, in some detail, focusing on convergent and discriminant aspects, such as the multitrait–multimethod approach, the nature of method variance, and multiple sources of data. Although one might quibble with some of Messick’s (1989) particular categories (e.g., some of them seem to overlap), we view his formulation as comprehensive and heuristically useful. Most important, we agree with Messick’s view that evidence concerning traditional issues of reliability are part of the construct validation program, namely under the heading of generalizability, and that evidence about dimensionality must be considered in the context of structural validity. In the following section, we reconsider these issues, now from the perspective of the measurement model in SEM.

MODEL TESTING IN CONSTRUCT VALIDATION AND SCALE CONSTRUCTION

The measurement model in SEM is based on confirmatory factor analysis (CFA); Loehlin (1998), McArdle (1996), and Bollen and Long (1993) provided recent introductions (see also Wegener & Fabrigar, this volume, Ch. 16). 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 briefly discuss a simple numerical example.

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

Like all factor analytic procedures (Ployd & Widaman, 1995; Tinsley & Tinsley, 1987), 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 be displayed graphically, allowing us to effectively communicate the various assumptions incorporated in each model. Some examples are shown in Figure 13.4. Figure 13.4a shows a common-factor model, in which a single underlying construct \( S \) (shown as a circle on the top) is assumed to give rise to the correlations among the six items or responses \( R_i \) to \( R_6 \) (the observed variables, shown in squares). Following established convention (Bentler, 1980), circles 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 \( R_i \) has two arrows leading to it. The arrow from the latent construct \( S \) is a factor 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 six observed variables covary only because they all measure the same underlying construct \( S \). In other words, we hypothesize that the only thing the items have in common is this 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 Panel a of Figure 13.4 with the one in Panel b that postulates two factors \( S_1 \) and \( S_2 \) influencing responses to six items. Here we are hypothesizing two distinct constructs, rather than one. Note that 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 two items are influenced by the first construct but not the second construct, whereas the
Figure 13.4. Measurement models in structural equation modeling: (a) a one-common-factor model, (b) a two-factor oblique model, and (c) a model showing one common factor related to a criterion.
last two items are influenced only by the second construct and not the first. In other words, these items can be uniquely assigned to only one construct, which much simplifies the measurement model. With two constructs in the measurement model, we can also address issues of discriminant validity. Whereas the item’s loading on the construct of interest represents convergent validity and its unique variance random error, its loading on constructs other than the intended one is relevant to discriminant validity.

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). At the level of the constructs, this correlation tells us about discriminant validity. If the correlation is very high (e.g., .90), we would worry that the two constructs are not distinguishable 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 could be components of a broader, superordinate construct that includes them both. These issues involve questions about the dimensionality and internal structure of the constructs being measured. We discussed these issues earlier in the section on reliability but, as we argued in the section on validity, dimensionality issues are part of the construct validation program (see Table 13.4) because they concern the structural validity of the interpretation of our measure.

**STRUCTURAL VALIDITY EXAMINED WITH SEM: AN EMPIRICAL EXAMPLE.** Structural validity issues resurface with great regularity in the social–personality literature. Some of the most popular constructs have endured protracted debates about their validity: self-monitoring, attributional style, hardiness, Type A coronary-prone behavior pattern, and most recently need for closure (Hull, Lehn, & Tedlie, 1991; Neuberg, Judice, & 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 instructive to consider how SEM approaches can help address the underlying issues and to provide a numerical though manageable example as an illustration.

For the purpose of this illustration we constructed two hypothetical scales and then used actual data from participants who had rated themselves on a number of personality-descriptive adjectives and phrases (Benet-Martínez & John, 1998); we used a large sample (N = 450) because small sample sizes can create problems for SEM estimation procedures (McArdle, 1996).

The first scale was intended as a measure of impulsivity (vs. inhibition), a construct of long-standing interest and debate (e.g., J. Block, 1995; Kagan & Snidman, 1991). To address content validation early on, we defined our universe of item content from the perspective of generalizability theory, using a design that varied two facets of generalizability: context (task vs. social) and construct pole (impulsive vs. inhibited). For the high (impulsive) pole, we selected from our existing item pool three items to represent task contexts (careless, disorganized, lazy) and two items to represent social contexts (enthusiastic and assertive); for the low (inhibited) pole of the construct, two items each for task contexts (persevering and thorough) and social contexts (reserved and shy).

The second scale, briefly, was intended as a measure of extraversion and sampled from previously studied content facets of the construct, namely talkativeness and self-assertion; again, it included both extraverted items (assertive, has an assertive personality, bold, verbal, is talkative, talkative) and introverted items (untalkative, tends to be quiet, and timid). To begin with, we assumed that each scale is unidimensional (of course, we had doubts about one of the scales, as will soon become clear).

What do we find when we apply the traditional analyses of internal consistency and exploratory factor analysis to these two scales? Table 13.5 summarizes the

<table>
<thead>
<tr>
<th></th>
<th>&quot;Impulsivity&quot; (two-dimensional)</th>
<th>Extraversion (one-dimensional)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha of total scale</td>
<td>.79</td>
<td>.90</td>
</tr>
<tr>
<td>First principal component</td>
<td>38%</td>
<td>58%</td>
</tr>
<tr>
<td>Second principal component</td>
<td>20%</td>
<td>13%</td>
</tr>
<tr>
<td>Correlation between subscales</td>
<td>-.30</td>
<td>.67</td>
</tr>
</tbody>
</table>

major results. As we cautioned earlier, one should not calculate alphas before the unidimensionality of a scale has been verified. Indeed, the alphas seem reasonably high for these relatively short scales, and most journal editors would consider even the lower alpha of .79 quite acceptable (in fact, the 25-item self-monitoring scale had a lower alpha than this 9-item scale; Snyder, 1987). Moreover, the item–total correlations were all substantial for all items on both scales.

What about the structural validity of these scales? Exploratory factor analyses resulted in eigenvalues (Table 13.5 shows the corresponding percentages of variance accounted for by the first and second unrotated principal component) that make it hard to tell conclusively if we need one or two factors, although for the impulsivity scale the evidence points to two factors. The loadings for two rotated factors, shown in Figures 13.5a and 13.5b, are more informative. For impulsivity, the items defining our four a priori facets nicely hang together, and the items show excellent simple structure. However, the exploratory factor solution seems inconsistent with a one-dimensional impulsivity theory because the social impulsivity items and the task impulsivity items formed two distinct factors.

To find out whether the social and task impulsivity factors have a substantial positive relation, as the general impulsivity account would suggest, we formed two scales and intercorrelated them. Social impulsivity (enthusiastic and assertive vs. reserved and shy) correlated .30 with task impulsivity (careless, lazy, and disorganized vs. thorough and persevering), thus failing to produce the predicted positive correlation—the facets do not hang together the way they should.

In contrast, if we rotate two factors for the extraversion scale (see Figure 13.5b), we find much less evidence for simple structure; instead, the variables all fall into a positive manifold (all items are either in the high-high or low-low quadrant), with the two a priori facets forming somewhat separable assertiveness and talkativeness clusters, especially within the upper-right quadrant. As expected from this loading pattern, the two clusters (when scored as scales) were correlated .67—a positive and substantial correlation (see Table 13.4). These exploratory analyses leave us with some alternative hypotheses that we can test against the a priori models, using SEM.

The SEM analyses are summarized very briefly in Table 13.6. 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 comparative fit indices show that we can clearly reject the one-factor model for the impulsivity scale; a test comparing it with the uncorrelated two-factor

**Figure 13.5.** Plot of exploratory factor loadings (after Varimax rotation) for (a) the items assumed to relate to impulsivity, and (b) items assumed to relate to Extraversion.
model shows that the latter provides a significantly better fit for the data. Not surprisingly, fit is improved further (and significantly) when correlations between the two factors are freely estimated. Note, however, that this model requires the estimation of another parameter, namely the -.30 correlation between the two factors.

These findings are inconsistent with the general impulsivity hypothesis: Because the two factors do not correlate positively, we cannot conclude that the scale consists of two impulsivity facets that together form the superordinate construct. Rather, we might conjecture that we are measuring abbreviated versions of the familiar Big Five factors of extraversion (defined by our social impulsivity items) and conscientiousness (defined by the task impulsivity items reversed-scored). Indeed, the present findings are quite consistent with numerous studies of much broader sets of personality descriptors (e.g., John, 1990; John & Srivastava, 1999).

CRITERIA FOR UNIDIMENSIONALITY. We can push the model comparison approach even farther. Rather than letting SEM estimate the correlation (or covariation) between the latent factors as a free parameter, we can fix it at a value that would allow us to make strong inferences about the independence of the two constructs. This is what we did in the last two models in Table 13.6: We set a specific decision rule, fixing the correlation to one of two decisive values and asking which one fit the data more closely.

Hattie (1985) emphasized that researchers need to formulate decision criteria that help them decide “how close a set of items is to being a unidimensional set” (p. 159). In most real data, just as in the present one, unidimensionality is not simply present or absent. Thus, we need to make a conceptual argument at what levels of intercorrelation we will call a measure relatively unidimensional or relatively multidimensional. Although reasonable people can disagree about any one cut-off point because it is inherently arbitrary, we are prepared to argue that factor intercorrelations as low as .20 would indicate relative independence, whereas correlations as high as .80 suggest such substantial overlap that a one-factor model should be preferred on the basis of parsimony.

As shown in Table 13.6, this decision rule allowed us to differentiate between the two models. The low-intercorrelation model provided a significantly better fit for the “impulsivity” items than did the high-intercorrelation model, thus correctly identifying this item set as measuring two essentially independent constructs; we say “correctly” here because we had in fact constructed this set by drawing items from uncorrelated Big Five self-report scales for conscientiousness and extraversion, respectively. In contrast, note that the high-intercorrelation model provided a significantly better fit for the extraversion scale. This result suggests that this item set is best interpreted as measuring an essentially unidimensional construct with two highly correlated item clusters, which might be interpreted as talkative and assertive manifestations of extraversion.

MORE COMPLEX MODELS INCLUDING EXTERNAL VALIDITY. In a fully developed construct validation program, of course, we would not stop here. For the “impulsivity” item set, we would move on to testing the relations of these two SEM-based constructs with other measures of extraversion and conscientiousness, preferably drawn from other data sources, such as peer ratings or behavioral observations. Using an MTMM design to address external validity, we would gather evidence both about convergent validity (e.g., self-reported conscientiousness with measures of conscientiousness drawn from other data sources) and discriminant validity (e.g., measures of conscientiousness with measures of extraversion from the same data source).

Again, we would use SEM procedures for these additional validation steps, as suggested by the simplified

<table>
<thead>
<tr>
<th>Table 13.6. Structural Validity Example in the SEM Approach: Summary of Models Tested and Their Fit Indices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model tested</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>One factor only</td>
</tr>
<tr>
<td>Two uncorrelated factors</td>
</tr>
<tr>
<td>Two correlated factors, r freely estimated</td>
</tr>
<tr>
<td>Two correlated factors, r set to .20</td>
</tr>
<tr>
<td>Two correlated factors, r set to .80</td>
</tr>
<tr>
<td>Two correlated factors, r set to .80</td>
</tr>
</tbody>
</table>

model in Figure 13.4.c. This model shows how we can represent a unidimensional measurement model for construct $S_4$ and a unidimensional criterion construct $S_4$, along with a predictive (or convergent) validity relation represented by the arrow from $S_4$ to $S_4$. 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 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, Figure 13.2) provided an elegant model for a more complete representation of the construct validation program.

Many readers might benefit from an example with more extensive numerical illustrations of SEM than we could provide here. We recommend an early paper by Judd, Jessar, and Donovan (1986), who examined the construct validity of a 9-item scale designed to measure attitudinal tolerance of deviance, including attitudes toward shoplifting, lying, and getting into fights. This construct postulates the existence of an underlying general attitude toward deviance manifested in self-reported attitudes about specific deviant behaviors. To elaborate four aspects of the construct validity of this measure, Judd et al. (1986) used various SEM procedures. First, to examine the convergent validity (or internal consistency) of the 9 items, they analyzed their intercorrelations (or covariances), testing hypotheses about structural validity (e.g., do all 9 items reflect a single common factor?). Second, to examine external (or criterion) validity, they tested whether the construct relates to other constructs in the theoretically consistent ways (e.g., do these attitudinal items predict deviant behavior?). Third, to address discriminant validity, they measured discriminant relations regarding religious attitudes in terms of both structural validity (e.g., are the attitude-toward-deviance items reliably different from religious attitude items?) and criterion validity (e.g., are these items better at predicting deviant behaviors than are the religious attitude items?). Fourth, they investigated particular aspects of substantive validity, namely temporal stability and prediction of behavior over time; these substantive predictions are important because attitudes, like other personality constructs, refer to individual differences that are assumed to be relatively stable and enduring over time, rather than transitory states of short duration (Chaplin, John, & Goldberg, 1988).

Issues in Questionnaire 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 psychological measures but also during each stage of their development. We now consider the somewhat specialized case of questionnaire (or scale) construction. The first questionnaires were developed in the early 1900s, and since the 1950s the construction of questionnaires began to proliferate (Goldberg, 1971). We will argue that questionnaire construction, like the development of any psychological measure, must be considered in the context of a program of construct validation. Historically, however, the construct validation approach was not articulated until the 1950s, and the consensus in its favor has been building slowly and quietly, mostly since the 1970s, and it is far from complete. Three distinct schools of thought preceded it and retain adherents even today.

EVALUATION, RATIONAL-INTUITIVE, AND INTERNAL. Three approaches to questionnaire construction emerged in the 1950s: each was inspired by one particular type of validity (see Table 13.4) and aimed to maximize that particular type of validity while ignoring others (for reviews, see Burisch, 1984, 1986). Given today's perspective favoring an integrated construct approach, the ideological fervor of these three camps strikes us almost like self-parodies.

The so-called external approach emphasized maximizing criterion validity, seemingly lost in "a single-minded bivariate search for items that correlate with a chosen criterion" (Tellegen, 1985, p. 685). Typically, externally oriented researchers would administer large and atheoretically assembled sets of questionnaire items to preselected criterion and control groups (e.g., patients hospitalized for depression vs. patients admitted for surgical procedures) and then determine empirically which items significantly differentiated the two groups. The items that successfully differentiated between the groups would be retained to form the resulting scale (e.g., for depression), regardless of the actual item content or broader theoretical considerations. The most famous products of the external approach are the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1943) for clinical populations and the California Psychological Inventory (CPI; Gough, 1987) for normally functioning adults. The continued popularity of these instruments, conceived in the 1950s, is testimony to the endurance of the
approach. Although the obsession with criterion validity still persists in some literatures (e.g., on marital interaction and satisfaction), the external approach largely fell out of favor, primarily because the subtle and theoretically opaque items did not form psychologically coherent and heuristic constructs, were hard to replicate, and required a rather inefficient scale-construction process.

At the other extreme of the dust bowl empiricists were those psychologists who had detailed theories they did not doubt. Thus they felt free to focus solely on the content and face validity of their measures. Various labeled the rational, intuitive, or deductive approach, they easily generated items on the basis of their theories. The resulting scales, face-valid with obvious item content, proved remarkably popular, if not always with other researchers then certainly with the test-taking public. In fact, this approach gave birth to the Myers-Briggs Type Indicator (MBTI; Myers & McCaulley, 1983), based partly on Carl Jung's type theories. Without much evidence for its external, structural, or substantive validity, the MBTI nonetheless became the most popular personality questionnaire in this country. To the eternal embarrassment of research psychologists, the MBTI continues to be used at major research universities in applied contexts, such as counseling and career advising. On the brighter side, the deductive approach eventually developed into the construct approach, which is, as we have described above, more interested in empirical evidence that might turn out to disconfirm one's favorite theory.

Finally, an emphasis on structural validity and the growing availability of exploratory factor analysis in the 1950s and 1960s gave rise to the internal or inductive approach to questionnaire construction. As the label suggests, the focus was on discovering the factor structure of large item sets, often assembled with little concern for particular content representation or selection. The early-factor-analytic personality models of Cattell, Eysenck, and Guilford were based on this approach and dominated until the mid-1980s when they gave way to the emerging consensus on the Big Five dimensions (Goldberg, 1993; John, 1990; John & Srivastava, 1999). The preoccupation with the internal factor structure in self-reports came at the expense of other sources of validity evidence. Partly because the dimensions emerging from factor analyses were assumed to be "real," theoretical construct definitions, substantive validity evidence, and criterion validation against measures of behaviors were deemed of secondary interest.

It is easily apparent that each of these three approaches, in its pure form, had a great strength that was also its greatest failure, namely its single-minded pursuit of just one type of validity evidence. Obviously, some kind of integration was needed. Although the conceptual foundations had been laid already in the 1950s, the construct validation approach emerged only gradually, as the three earlier approaches grew softer around the edges and eventually became indistinguishable.

**RECAPITULATION: MODERN CONSTRUCT-ORIENTED SCALE CONSTRUCTION.** Few, if any, scales or measures today are constructed according to just one of the early approaches. Most researchers 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, in considerable detail, 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 13.4 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 (a) generating hypotheses; (b) building a model and plausible alternatives; (c) generating items using construct definitions, generalizability facets, and content validation procedures as guides (for information about item and response formats, see Visser, Krosnick, & Lavrikas, this volume, Ch. 9); (d) gathering and analyzing data; (e) confirming and disconfirming the initial models; and (f) generating alternative hypotheses leading to improved models, additional and more content-valid items, 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, at least for a while, given the constraints and limits of real-life research.

There are a few admirable individuals who have completed what amounts to a lifelong and thus exhaustive (and exhausting) program of construct validation. Developing and validating a sentence completion test to measure ego development would seem a hopeless undertaking to most. Yet, Jane Loewinger, one of the field's premier psychometricians and pioneer of construct validation in the 1950s (e.g., Loewinger, 1957), took on this project and devoted a good 30 years of work to it, with impressive results (Loewinger, 1998a).
However, eventually she came to view her work on that sentence completion test as "completing a life sentence" (Loevinger, 1998b, p. 347).

Other examples include Robert Altemeyer's (1988) tireless efforts to sift the real construct of authoritarianism from the ashes of the F scale; after some 20-plus years of validation studies, and probably an equal number of scale reconstructions, his right-wing authoritarianism scale is now firmly embedded in a nomological network of generalizable construct relations. Douglas Jackson's (1971, 1984) construction and validation of the Personality Research Form (PRF) offers a more circumscribed and manageable example of a construct validation program that may also serve as a blueprint for other efforts. Jackson emphasized four broad steps: (a) substantive definitions of scale content, (b) sequential strategy in scale construction, (c) appraisal of the structural component of validity (including the suppression of unwanted response style variance), and (d) evaluation of the external component of validity. At each of these steps, Jackson also tried to foster the goals of scale generalizability and convergent and discriminant validity.

It is easily apparent that these four steps and two key concerns closely map onto our discussion of construct validation. In other words, scale construction and construct validation go hand in hand, and one cannot be separated from the other. Our final topic is a unique special set of measurement issues that scale construction efforts need to address, namely the language and culture contexts within which the measure is developed and used.

**CULTURAL AND TRANSLATION ISSUES IN QUESTIONNAIRE CONSTRUCTION.** After years of relative neglect, interest in cross-cultural research has been growing in the 1990s (Van de Vijver & Leung, 1997). There are both theoretical and practical reasons for examining psychological measures in cultures other than the United States. First, cross-cultural studies are needed to test the generalizability of our psychological theories and models. Second, given the increasing multiculturalism in the United States, cultural research is necessary to understand the psychological reality of cultural and ethnic minorities.

Methodological considerations are very important in cross-cultural research. Consider a researcher who wants to test a theory that two culture groups differ on a measure or that they show different correlates with a measure. First, the researcher must demonstrate that the same characteristic has been measured in the same way across the two groups. The most common research strategy has been to translate an original U.S.-developed measure to assess the construct of interest in a new culture. This imposed-etic strategy (Berry, 1980) is economical and efficient when we want to examine how a particular measure generalizes to other cultures. However, when we want to identify culture-specific aspects of a construct, the imposed-etic approach has serious limitations; using translated measures simply assumes that the construct is universal, thus ignoring meanings and indicators of the construct that are potentially culture-specific (Church & Katigbak, 1988).

The question whether imposed-etic measures overlook important domains of the local culture is at the core of the longstanding emic-etic debate (Berry, Poortinga, Segall, & Dasen, 1992), which contrasts the supposedly interculturally comparable, universal (etic) aspects of a construct with its culture-specific, indigenous (emic) aspects (Berry, 1980). On the one hand, an imposed-etic strategy is useful in that it makes cross-cultural comparisons feasible (i.e., statements about the similarity of two cultures require dimensional equivalence), yet its use may distort the meaning of the construct. On the other hand, a fully emic strategy is well-suited to identify culture-specific aspects of a construct (i.e., it is ecologically valid), but it renders comparisons across cultures virtually impossible (Berry, 1980). Note that the emic-etic distinction is not "either-or" but a matter of degree. Overlap between measures taken in different cultures is not simply present or absent, but rather varies in strength and breadth (Berry et al., 1992).

The current view is that emic and etic approaches render two distinct (though related) types of information. Thus, the two approaches need to be combined to provide a complete picture of cultural specificity and overlap (Benet-Martínez & Waller, 1997; Church & Katigbak, 1988; Yang & Bond, 1990). The use of a combined emic-etic approach requires the researcher (a) to identify the emic (indigenous) elements of the construct (through focus groups, interviews, or content analyses of popular media), and develop and administer measures that adequately tap these constructs; (b) to administer translated measures tapping imposed-etic constructs; and (c) to assess the specificity and overlap between imported and indigenous measures. By comparing the information yielded by imposed-etic and emic measures, the researcher can assess how well imported and indigenous constructs correspond and identify indigenous elements not represented by the imported (translated) instrument (for an illustration of this procedure, see Benet-Martínez & Waller, 1997).
An indispensable requirement for valid cross-cultural comparisons is conceptual equivalence, that is, symmetry in the meaning of different-language versions of a measure (see Van de Vijver & Leung, 1997, for a discussion of construct, measurement, and scalar equivalence in cross-cultural comparisons). One way to foster conceptual equivalence is to use the back-translation procedure (Brislin, 1980). One fluent bilingual (ideally an expert on the construct of interest) translates the instrument from the original language into the language of interest. A second bilingual expert independently translates these materials back into the original language. The combination of (a) comparing the back-translated version to the original, (b) discussions between translators, and (c) back-and-forth translations should lead to a final set of translated items that are symmetrically translatable to the original language counterparts.

Following careful back-translation procedures, construct validation procedures must be used to check the success of the translation. In comparisons of two monolingual samples (e.g., one Spanish speaking, the other English speaking), discrepancies in item and scale statistics indicate lack of equivalence but fail to reveal its source—it might be due to poor translations but sample and culture differences could also play a role. Thus, ideally, the two language versions are compared across samples of both monolinguals and bilinguals (see Benet-Martínez & John, 1998, for an illustration of how a bilingual design can be used to disentangle these confounds). Most recently, IRT (see our earlier discussion) has become an effective and popular tool for examining cross-cultural and cross-linguistic measurement invariance (e.g., Ellis, Becker, & Kimmel, 1993). If sample sizes are large enough, measurement invariance across languages can also be tested with CFA (Benet-Martínez & John, 1998). Together, CFA and IRT hold much promise to help resolve these special measurement problems in cross-cultural research (see Reise, Widaman, & Pugh, 1993).

CONCLUSIONS AND RECOMMENDATIONS

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 in social-personality research. We reviewed the traditional reliability coefficients but urged the reader to think about facets of generalizability, such as time, items, and observers, and to explicitly adopt a generalizability framework. We rallied against some of our pet peeves, such as the indiscriminate use of alpha, pointing out its limitations and arguing for more complex interpretations of this ubiquitous index. We discussed a unified conception of construct validity, suggesting that systematic construct validation efforts are needed to develop a theoretical understanding of our methods; this goal is worth a sustained program of research, rather than a few isolated criterion correlations sprinkled throughout the literature. We noted the voracious appetite our field has for “fast data” (the so easily obtained self-reports) and argued for a more diversified diet, calling for multimethod investigations as a rule, rather than the rare exception. We illustrated, briefly, the power of the no-longer new SEM techniques to address measurement problems, calling for their routine use, at least in samples of large size (of which we would like to see more, too).

This chapter has noted shortcomings in the current practice of measurement that one could deplore, and practices that ought to be changed. Nonetheless, we are upbeat about the future of measurement in social-personality psychology. In writing this chapter, we became particularly persuaded by the simple logic of comparative model-testing: We now see it as the best strategy for evaluating and improving our measurement procedures. We are confident that even though our journals still practice the archaic preoccupation with significance tests, comparative model-testing will catch on, eventually, and so will the powerful tools provided by SEM. Of course, it won’t happen tomorrow. As Jacob Cohen (1990) concluded from his 40 years of research on methodology, the “inertia” of methodological advance is enormous “but I do not despair . . . these things take time” (p. 1311).

REFERENCES


Tellegen, A. (1985). Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on
Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the Absorp-