CHAPTER 9

Validity

Estimating and Evaluating Convergent and Discriminant Validity Evidence

The previous chapter presented conceptual perspectives on validity, and it summarized the types of evidence used to gauge construct validity. As described in that chapter, convergent and discriminant evidence reflects the degree to which test scores have the "correct" patterns of associations with other variables. In this chapter, we focus more deeply on the way in which convergent and discriminant evidence can be evaluated, and we discuss issues bearing on the interpretation of convergent and discriminant evidence.

To restate the issue briefly, psychological constructs are embedded in a theoretical context. That is, the conceptual foundation of a construct includes the connections between the construct and a variety of other psychological constructs. The interconnections between a construct and other related constructs have been called a *nomological network,* which refers to the network of "meaning" surrounding a construct (Cronbach & Meehl, 1955). For example, Baumeister and Leary (1995) introduced the concept of a "need to belong," which was defined as the "drive to form and maintain at least a minimum quantity of lasting, positive, and significant interpersonal relationships" (p. 497). Although they hypothesized that this is a fundamental human need, they also observed that people differ in the degree to which they experience this need. Some people have a relatively great need to experience frequent interactions within close and caring relationships, while other people seem to need such interactions less. Leary, Kelly, Cottrell, and Schreindorfer (2006) theorized about the nomological network surrounding the need to belong. Specifically, they suggested that the need to belong was somewhat similar to characteristics such as the need for affiliation, the need for intimacy, sociability, and extraversion. Furthermore, the need to belong was essentially unrelated to constructs such as conscientiousness, openness to experience, and self-esteem.

The nomological network of associations among constructs dictates a particular pattern of associations among measures of those constructs. The nomological network surrounding a construct suggests that a measure of the construct should be strongly associated with measures of some constructs but weakly correlated with measures of other constructs. For example, Leary et al. (2006) predicted that their 10-item "Need to Belong" (NTB) measure would be weakly to moderately correlated with measures of need for affiliation, sociability, and extraversion; negatively correlated with a measure of social isolation; and essentially uncorrelated with measures of conscientiousness, openness to experience, and self-esteem. Their predictions guided their evaluation of the convergent and discriminant quality of the NTB.

A crucial part of the validation process is estimating the degree to which test scores actually show the predicted pattern of associations. In this chapter, we present some methods used in this process, some important factors affecting the outcome of the process, and some key considerations in interpreting the outcomes.

Methods for Evaluating Convergent and Discriminant Validity

There are at least four methods used to evaluate the degree to which measures show convergent and discriminate associations. These procedures differ in several ways—some are more conceptually complex than others, some can be more statistically complex than others, some are decades old while others are relatively new, and some require more explicit predictions than others. Despite these differences, the following methods are (or might become) common and useful ways of evaluating convergent and discriminant validity evidence.

Focused Associations

Some measures have clear relevance for a few very specific variables. Evaluating the validity of interpretations for such measures can focus on the associations between test scores and those relatively few specific variables. In a sense, these specific associations are "make-or-break" in terms of the convergent and discriminant validity evidence for such measures. Research verifying those crucial predicted associations provides strong validity evidence, but research failing to verify the associations casts serious doubt on validity.

As mentioned in the previous chapter, the Scholastic Assessment Test (SAT) Reasoning test is intended to be "a measure of the critical thinking skills [needed] for academic success in college" (College Board, 2006). This description implies that two kinds of variables might be particularly critical for evaluating the SAT Reasoning test. First, as a measure of "critical thinking skills," the SAT should be associated with other measures of relevant critical thinking skills. Second, because it is intended to assess a construct required for "academic success in college," the SAT should be associated with measures of collegiate academic performance.

In establishing the quality of the SAT, the College Board appears to be most concerned with the latter issue. A number of documents that are made available to students, educators, and prospective researchers emphasize the correlation between SAT scores and academic indicators such as first-year college grades. For example, the *SAT Program Handbook,* published by the College Board for school counselors and admissions officers, includes several references to validity (College Board, 2006). In the first section regarding validity, the *Handbook* states that a study of more than 110,000 students from more than 25 colleges revealed an average correlation of .55 between SAT scores and freshman grades. The handbook goes on to mention additional studies providing predictive validity evidence for the SAT in relation to college grades. Clearly, the College Board focuses its validity argument on the correlations between the SAT and a specific set of criterion variables related to academic performance in college.

Thus, one method for evaluating the validity of test interpretations is to focus on a few highly relevant criterion variables. To the degree that test scores are indeed correlated with those crucial variables, test developers and test users gain increased confidence in the test. Those correlations, sometimes called *validity coefficients,* are fundamental for establishing validity. If research reveals that a test's validity coefficients are generally large, then test developers, users, and evaluators will have increased confidence in the quality of the test as a measure of its intended construct.

Validity generalization is a process of evaluating a test's validity coefficients across a large set of studies (F. L. Schmidt, 1988; F. L. Schmidt & Hunter, 1977). Unlike the SAT, many measures used in the behavioral sciences rely on validity evidence obtained from relatively small studies. In fact, many if not most validity studies include fewer than 400 participants—particularly if those studies include anything besides self-report data. Often a researcher conducting a single validity study will recruit a sample of 50 to 400 participants, administer the measure of interest to those participants, assess additional criterion variables deemed relevant, and compute the correlation between the scores on the measure of interest and scores on the criterion measures. Such studies are the basis of many measures used for research in personality psychology, clinical psychology, developmental psychology, social psychology, organizational psychology, and educational psychology. Individual validity studies often include relatively small samples due to limits on researchers' time, funding, and other resources.

Although studies with relatively small samples are common and are conducted for many practical reasons, they do have a potentially important drawback. A study conducted at one location with one type of population might provide acceptable convergent and discriminant validity evidence for a measure, but the results might not generalize to another location or another type of population.

For example, the results of a study of bank employees might demonstrate that scores on the Revised NEO Personality Inventory (NEO-PI-R) Conscientiousness scale are relatively good predictors of job performance for bank tellers. Although this is potentially valuable and useful evidence for human resources directors in the banking industry, do these results offer any insight for human resources directors in the accounting industry, the real estate industry, or the sales industry? That is, is the association between conscientiousness scores and job performance strong only for bank tellers, or does it generalize to other groups? Perhaps the trait of conscientiousness is more relevant for some kinds of jobs than for others. If so, then we should not assume that the NEO-PI-R is not a valid predictor of job performance in all professions.

Validity generalization studies are intended to evaluate the predictive utility of a test's scores across a range of settings, times, situations, and so on. A validity generalization study is a form of meta-analysis, which combines the results of several smaller individual studies into one large analysis (F. L. Schmidt, Hunter, Pearlman, & Hirsh, 1985). For example, we might find 25 studies examining the association between the NEO-PI-R Conscientiousness scale and job performance. One of these studies might have examined the association among bank tellers, another might have examined the association within a sample of school teachers, another might have examined the association within a sample of salespersons, and so on. Each study might include a different kind of profession, but each study also might include a different way of measuring job performance. For instance, some studies might have relied on managers' ratings of employees' job performance, while other studies might have used more concrete measures of job performance such as "dollars sold." Thus, we might find that the 25 different studies reveal apparently different results regarding the strength of association between NEO-PI-R Conscientiousness scores and job performance.

Validity generalization studies can address at least three important issues. First, they can reveal the general level of predictive validity across all of the smaller individual studies. For example, the analysis of all 25 studies in our conscientiousness example might reveal that the average validity correlation between NEO-PI-R Conscientiousness scores and job performance is .30. Second, validity generalization studies can reveal the degree of variability among the smaller individual studies. We might find that, among the 25 studies in our generalization study, some have quite strong associations between NEO-PI-R Conscientiousness scores and job performance (say, correlations of .40 to .50), while others have much weaker associations (say, correlations of .00 to .10). If we found this kind of variability, then we might need to conclude that the association between NEO-PI-R Conscientiousness scores and job performance does not generalize across the studies. Conversely, our validity generalization study might reveal that, among the 25 studies in our generalization study, almost all have moderate associations between NEO-PI-R Conscientiousness scores and job performance (say, correlations of .20 to .40). If we found this smaller amount of variability among the 25 studies, then we might conclude that the association between NEO-PI-R Conscientiousness scores and job performance does in fact generalize across the studies quite well. Either way, the finding would be important information in evaluating the validity and use of the NEO-PI-R in hiring decisions.

The third issue that can be addressed by validity generalization studies is the source of variability among studies. If initial analyses reveal a wide range of validity coefficients among the individual studies, then further analyses might explain why the studies' results differ from each other. For example, we might find a methodological difference among studies that corresponds to the validity coefficient differences among the studies. We might discover that strong validity coefficients were found in studies in which managers provided ratings of job performance but that weaker validity coefficients were found in studies in which concrete measures such as "dollars sold" were used to assess job performance. Thus, differences in the measurement of the criterion variable seem to contribute to differences in the size of the validity coefficient. This kind of methodological source of variability should be considered when evaluating the implications of the general level and variability of validity coefficients across studies.

In sum, some psychological tests are expected to be strongly relevant to a few specific variables. If research confirms that such a test is indeed strongly associated with its specific criterion variables, then test developers, users, and evaluators gain confidence that the test scores have good convergent validity as a measure of the intended construct. A validity generalization study evaluates the degree to which the association between a test and an important criterion variable generalizes across individual studies that cover a range of populations, settings, and so on.

Sets of Correlations

The nomological network surrounding a construct does not always focus on a small set of extremely relevant criterion variables. Sometimes, a construct's nomological network touches on a wide variety of other constructs with differing levels of association to the main construct. In such cases, researchers evaluating convergent and discriminant validity evidence must examine a wide range of criterion variables.

In such cases, researchers often compute the correlations between the test of interest and measures of the many criterion variables. They will then "eyeball" the correlations and make a somewhat subjective judgment about the degree to which the correlations match what would be expected on the basis of the nomological network surrounding the construct of interest.

For example, Hill et al. (2004) developed a new measure of perfectionism, and they presented evidence of its convergent and discriminant validity. The Perfectionism Inventory (PI) was designed to measure eight facets of perfectionism, so it was intended to have a multidimensional structure (see the discussion on internal structure in the previous chapter). Specifically, the PI was designed to assess facets such as concern over mistakes, organization, planfulness, striving for excellence, and need for approval. To evaluate the convergent and discriminant validity evidence, participants were asked to complete the PI along with measures of 23 criterion variables. Criterion variables included other measures of perfectionism. In addition, because perfectionism is associated with various kinds of psychological distress, the criterion variables included measures of several psychological symptoms (e.g., obsessive-compulsive disorder, anxiety, fear of negative evaluation). The correlations between the PI scales and the 23 criterion scales were presented in a correlation matrix that included more than 200 correlations (see Table 9.1).

To evaluate the convergent and discriminant validity evidence, Hill and his colleagues (2004) carefully examined the correlations and interpreted them in terms of their conceptual logic. For example, they noted that the Concern Over Mistakes

Table 9.1 Example of Sets of Correlations in the Validation of the Perfectionism Inventory

Correlations Between Perfectionism Indicator Scales and Related Measures

Scale	*CM*	*HS*	*NA*	*OR*	*PP*	*L*	*RU*	*SE*	*CP*	*SEP*	*PI–C*
Perfectionism: MPS–F[a]											
Concern Over Mistakes	.82	.43	.58	.18	.38	.30	.70	.52	.47	.78	.72
Doubts About Actions	.63	.37	.60	.24	.20	.38	.70	.43	.47	.67	.65
Parental Criticism	.41	.25	.20	–.03^{ns}	.60	.02^{ns}	.32	.17	.14	.49	.36
Parental Expectations	.31	.27	.18	.07^{ns}	.85	.06^{ns}	.29	.32	.23	.53	.43
Personal Standards	.47	.50	.36	.45	.3	.44	.52	.72	.70	.55	.71
Organization	.12	.36	.18	.89	.11**	.49	.31	.51	.76	.23	.55
Perfectionism: MPS–HF[b]											
Self-Oriented	.47	.42	.34	.47	.42	.45	.55	.79	.71	.57	.73
Other-Oriented	.33	.62	.14**	.29	.30	.26	.37	.42	.53	.36	.51
Socially Prescribed	.65	.35	.49	.16**	.58	.21	.61	.42	.38	.74	.65
Symptoms: BSI[c]											
Somatic Complaints	.35	.14*	.31	.13*	.11*	.13*	.34	.17	.19	.35	.31
Depression	.46	.16**	.46	.03^{ns}	.15**	.18	.46	.13*	.17	.49	.39
Obessive–Compulsive	.40	.14**	.46	.08^{ns}	.10**	.19	.46	.18	.19	.45	.37
Anxiety	.42	.28	.42	.22	.25	.25	.49	.29	.35	.50	.49
Interpersonal Sensitivity	.52	.18	.68	.17	.13*	.22	.56	.27	.28	.60	.51

Correlations Between Perfectionism Indicator Scales and Related Measures

Scale	*CM*	*HS*	*NA*	*OR*	*PP*	*L*	*RU*	*SE*	*CP*	*SEP*	*PI–C*
Hostility	.41	.30	.31	.10*	.21	.05^{ns}	.39	.15**	.20	.42	.36
Phobic Anxiety	.39	.14**	.39	.13*	.15**	.13*	.39	.15**	.21	.42	.37
Paranoia	.48	.28	.49	.18	.21	.21	.54	.30	.33	.55	.51
Psychoticism	.49	.19	.48	.09^{ns}	.16**	.19	.49	.17	.22	.51	.43
Global Severity Index	.54	.24	.55	.16	.20	.21	.57	.25	.29	.59	.51
Obsessive-Compulsive Inventory^{d}											
Frequency	.43	.24	.45	.39	.08^{ns}	.34	.52	.42	.47	.47	.54
Distress	.50	.28	.49	.40	.03^{ns}	.33	.60	.44	.48	.51	.57
Fear of Negative Evaluation^{a}	.63	.26	.83	.16	.20	.31	.64	.33	.34	.73	.62
Social Desirability: MCSDS^{c}	–.15**	–.17	–.09*	–.04^{ns}	–.14**	–.09*	–.18	–.16	–.12**	–.18	–.18

Notes: For all correlations, $p < .001$ (except as noted). CM = Concern Over Mistakes; HS = High Standards for Others; NA = Need for Approval; OR = Organization; PP = Perceived Parental Pressure; PL = Planfulness; RU = Rumination; SE = Striving for Excellence; CP = Conscientious Perfectionism; SEP = Self-Evaluative Perfectionism; PI-C = Perfectionism Indicator Composite score; MPS-F = Frost's Multidimensional Perfectionism Scale; MPS-HF = Hewittt and Flett's Multidimensional Perfectionism Scale; BSI = Brief Symptom Index; MCSDS = Marlowe-Crowne Social Desirability Scale.

$^{a}n = 613$. $^{b}n = 355$. $^{c}n = 368$. $^{d}n = 207$.

$^{*}p < .05$, one-tailed. $^{**}p < .01$, one-tailed. $^{ns}p > .05$, all one-tailed.

Source: Copyright 2004 from *Journal of Personality Assessment, 92*(1), 80–91 by Hill, et al. Reproduced by permission of Taylor & Francis Group, LLC., http://www.taylorandfrancis.com.

scale of the PI was strongly associated with a Concern Over Mistakes scale from a different measure of perfectionism. Similarly, they noted that the Striving for Excellence scale of the PI was strongly associated with a Personal Standards scale (i.e., indicating high expectations for one's performance and an inclination to base self-appraisal on performance) and a Self-Oriented Perfectionism scale (i.e., indicating unrealistic standards for performance and the tendency to fixate on imperfections in one's performance) from other measures of perfectionism. They also examined the associations between the PI scales and the various measures of psychological distress. For example, they noted that three of the PI scales—Rumination, Concern Over Mistakes, and Need for Approval—were strongly associated with fear of negative evaluation and with the frequency and severity of symptoms of obsessive-compulsive disorder.

This approach to evaluating validity is common. Researchers gather a large amount of data concerning the test of interest and measures from a variety of other tests. They then examine the pattern of correlations, and they judge the degree to which the pattern "makes sense" given the conceptual meaning of the construct being assessed by the test.

Multitrait-Multimethod Matrices

One of the most influential papers in the history of psychological measurement was a paper published in 1959 by Campbell and Fiske. In this paper, Campbell and Fiske built upon the concept of construct validity as articulated by Cronbach and Meehl (1955). As we have already discussed, Cronbach and Meehl outlined a conceptual meaning of construct validity based on the notion of a nomological network. Although their paper was a hugely important conceptual advance, Cronbach and Meehl did not present a way to evaluate construct validity in a rigorous statistical manner. Campbell and Fiske developed the logic of a multitrait-multimethod matrix (MTMMM) as a statistical and methodological expansion of the conceptual work done by Cronbach and Meehl.

For the analysis of an MTMMM, researchers evaluating construct validity obtain measures of several traits, each of which is measured through several methods. For example, researchers evaluating a new self-report questionnaire of social skill might ask participants to complete the questionnaire along with self-report measures of several other traits such as impulsivity, conscientiousness, and emotional stability. In addition, they might ask close acquaintances of the participants to provide ratings of the participants' social skill, impulsivity, conscientiousness, and emotional stability. Finally, they might hire psychology students to interview each participant and then provide ratings of the participants' social skill, impulsivity, conscientiousness, and emotional stability. Thus, for each participant, the researchers obtain data relevant to multiple traits (social skill, impulsivity, conscientiousness, and emotional stability), each of which is measured through multiple methods (self-report, acquaintance ratings, and interviewer ratings).

The overarching purpose of the MTMMM analysis is to set clear guidelines for evaluating convergent and discriminant validity evidence. This purpose is partially served through evaluating two importantly different sources of variance that might affect the correlations between two measures—trait variance and method variance. To understand these sources of variance, imagine that researchers examining the new self-report measure of social skill find that scores on their measure are highly correlated with scores on a self-report measure of emotional stability. What does this finding tell them?

Strictly speaking, the finding tells them that people who say that they are relatively socially skilled tend to say that they are relatively emotionally stable. But does this finding reflect a purely psychological phenomenon in terms of the associations between two constructs, or does it reflect a more methodological phenomenon separate from the two constructs? In terms of psychological phenomena, it might indicate that the trait of social skill shares something in common with the trait of emotional stability. That is, the measures might share trait variance. For example, people who are socially skilled might tend to become emotionally stable (perhaps because their social skill allows them to create social relationships that have emotional benefits). Or people who are emotionally stable might tend to become more socially skilled (perhaps because their stability allows them to be comfortable and effective in social situations). Or it might be that social skill and emotional stability are both caused by some other variable altogether (perhaps there is a genetic basis that influences both stability and social skill). Each of these explanations indicates that the two traits being assessed—social skill and emotional stability—truly overlap in some way. Because the traits share some commonality, the measures of those traits are correlated with each other.

Despite our inclination to make a psychological interpretation of the correlation between social skill and emotional stability, the result might actually have a relatively nonpsychological basis. Recall that our example was based on the correlation between self-report measures. Thus, the correlation might be produced simply by *shared method variance.* That is, the correlation is positive because it is based on two measures derived from the same source—respondents' self-reports in this case. When measures are based on the same data source, they might share properties aside from the main constructs being assessed by the measures.

For example, people might tend to see themselves in very generalized terms—either in generally "good" ways or in generally "bad" ways. Therefore, a positive correlation between self-reported social skill and self-reported emotional stability might be due solely to the fact that people who report high levels of social skill simply tend to see themselves in generally good ways; therefore, they also tend to report high levels of emotional stability. Similarly, people who report low levels of social skill simply tend to see themselves in generally bad ways; therefore, they also tend to report low levels of emotional stability. In this case, the apparent correlation between social skill and emotional stability does not reflect a commonality between the two traits being assessed by the measures. Instead, the correlation is simply a by-product of a bias inherent in the self-report method of measurement. That is, the correlation is an "artifact" of the fact that the two measure share the same method (i.e., self-report). Testing experts would say that the ratings share method variance.

Due to the potential influences of trait variance and method variance, a correlation between two measures is a somewhat ambiguous finding. On one hand, a strong correlation (positive or negative) could indicate that the two measures share trait variance—the constructs that they are intended to measure have some commonality. On the other hand, a strong correlation (again, positive or negative) could indicate that the two measures share method variance—they are correlated mainly because they are based on the same method of measurement.

The ambiguity inherent in a correlation between two measures cuts both ways; it also complicates the interpretation of a *weak* correlation. A relatively weak correlation between two measures could indicate that the measures do not share trait variance—the constructs that they are intended to measure do not have any commonality. However, the weak correlation between measures could reflect differential method variance, thereby masking a true correlation between the traits that they are intended to assess. That is, the two traits actually could be associated with each other, but if one trait is assessed through one method (e.g., self-report) and the other is assessed through a different method (e.g., acquaintance report), then the resulting correlation might be fairly weak.

These ambiguities can create confusion when evaluating construct validity. Specifically, the influences of trait variance and method variance complicate the interpretation of a set of correlations as reflecting convergent and discriminant validity evidence. Each correlation represents a potential blend of trait variance and method variance. Because researchers examining construct validity do not know the true influences of trait variance and method variance, they must examine their complete set of correlations carefully. A careful examination can provide insight into trait variance, method variance, and, ultimately, construct validity. The MTMMM approach was designed to articulate these complexities, to organize the relevant information, and to guide researchers through the interpretations.

As articulated by Campbell and Fiske (1959), an MTMMM examination should be guided by attention to various kinds of correlations that represent varying blends of trait and method variance. Recall from our example that the researchers evaluating the new measure of social skill gathered data relevant to four traits, each which was measured through three methods. Let us focus on two correlations for a moment: (a) the correlation between the self-report measure of social skill and the acquaintance report measure of social skill and (b) the correlation between the self-report measure of social skill and the self-report measure of emotional stability. If the new self-report measure can be interpreted validly as a measure of social skill, then which of the two correlations should be stronger?

Based purely on a consideration of the constructs being measured, the researchers might predict that the first correlation will be stronger than the second. They might expect the first correlation to be quite strong—after all, it is based on two measures of the same construct. In contrast, they might expect the second correlation to be relatively weak—after all, social skill and emotional stability are different constructs. However, these predictions ignore the potential influence of method variance.

Taking method variance into account, the researchers might reevaluate their prediction. Note that the first correlation is based on two different methods of assessment, but the second correlation is based on a single method (i.e., two self-report measures). Thus, based on a consideration of method variance, the researchers might expect to find that the first correlation is weaker than the second.

As this example hopefully begins to illustrate, we can identify different types of correlations, with each type representing a blend of trait variance and method variance. Campbell and Fiske (1959) point to four types of correlations derived from an MTMMM (see Table 9.2).

Table 9.2 MTMMM Basics: Types of Correlations, Trait Variance, and Method Variance

Association Between the Two Constructs		*Method Used to Measure the Two Constructs*	
		Different Methods (e.g., Self-Report for One Construct and Acquaintance Report for the Other)	*Same Method (e.g., Self-Report Used for Both Constructs)*
Different constructs (not associated)	Label	Heterotrait-heteromethod correlations	Heterotrait-monomethod correlations
	Sources of variance	Nonshared trait variance and nonshared method variance	Nonshared trait variance and shared method variance
	Example	Self-report measure of social skill correlated with acquaintance report measure of emotional stability	Self-report measure of social skill correlated with self-report measure of emotional stability
	Expected correlation	Weakest	Moderate?
Same (or similar) constructs (associated)	Label	Monotrait-heteromethod correlations	Monotrait-monomethod correlations
	Sources of variance	Shared trait variance and nonshared method variance	Shared trait variance and shared method variance
	Example	Self-report measure of social skill correlated with acquaintance report measure of social skill	Self-report measure of social skill correlated with self-report measure of social skill (i.e., reliability)
	Expected correlation	Moderate?	Strongest

- *Heterotrait-heteromethod correlations* are based on measures of different constructs measured through different methods (e.g., a self-report measure of social skill correlated with an acquaintance report measure of emotional stability).
- *Heterotrait-monomethod correlations* are based on measures of different constructs measured through the same method (e.g., a self-report measure of social skill correlated with a self-report measure of emotional stability).
- *Monotrait-heteromethod correlations* are based on measures of the same construct measured through the different methods (e.g., a self-report measure of social skill correlated with an acquaintance report measure of social skill).
- *Monotrait-monomethod correlations* are based on measures of the same construct measured through the same method (e.g., a self-report measure of social skill correlated with itself). These correlations reflect reliability—the correlation of a measure with itself.

Campbell and Fiske (1959) articulated the definitions and logic of these four types of correlations, and they tied them to construct validity. A full multitrait-multimethod matrix of hypothetical correlations is presented in Table 9.3. The matrix includes 66 correlations among the three measures of four traits, along with 12 reliability estimates along the main diagonal. Each of these 78 values can be characterized in terms of the four types of correlations just outlined. The evaluation of construct validity, trait variance, and method variance proceeds by focusing on various types of correlations as organized in the MTMMM.

Evidence of convergent validity is represented by monotrait-heteromethod correlations, which are printed in boldface in the MTMMM. These values represent the correlations between different ways of measuring of the same traits. For example, the correlation between self-report social skill and acquaintance report social skill is .40, and the correlation between self-report social skill and interviewer report social skill is .34. These correlations suggest that people who describe themselves as relatively socially skilled (on the new self-report measure) tend to be described by their acquaintances and by the interviewers as relatively socially skilled. Monotrait-heteromethod correlations that are fairly strong begin to provide convergent evidence for the new self-report measure of social skill. However, they must be interpreted in the context of the other correlations in the MTMMM.

To provide strong evidence of its convergent and discriminant validity, the self-report measure of social skill should be more highly correlated with other measures of social skill than with any other measures. The MTMMM in Table 9.3 shows that, as would be expected, the monotrait-heteromethod correlations are generally larger than the heterotrait-heteromethod correlations, reflecting associations between measures of different constructs assessed through different methods (inside the dashed-line triangles). For example, the correlation between the self-report measure of social skill and the acquaintance report measure of emotional stability is only .20, and the correlation between the self-report measure of social skill and the interviewer report measure of conscientiousness is only .09. These correlations, as well as most of the other heterotrait-heteromethod correlations, are noticeably lower than the monotrait-heteromethod correlations discussed in the previous

Table 9.3 Example of MTMMM Correlations

		Self-Report				*Acquaintance Report*				*Interviewer Report*			
Methods	*Traits*	*Social Skill*	*Impulsivity*	*Conscien-tiousness*	*Emotional Stability*	*Social Skill*	*Impulsivity*	*Conscien-tiousness*	*Emotional Stability*	*Social Skill*	*Impulsivity*	*Conscien-tiousness*	*Emotional Stability*
Self-report	Social skill	(.85)											
	Impulsivity	.14	(.81)										
	Conscientiousness	.20	.22	(.75)									
	Emotional stability	.35	.24	.19	(.82)								
Acquaintance	Social skill	**.40**	.14	.10	.22	(.76)							
	Impulsivity	.13	**.32**	.13	.19	.18	(.80)						
	Conscientiousness	.09	.17	**.36**	.14	.14	.26	(.68)					
	Emotional stability	.20	.23	.11	**.41**	.30	.28	.18	(.78)				
Interviewer report	Social skill	**.34**	.11	.19	.20	**.23**	.01	.11	.19	(.81)			
	Impulsivity	.03	**.25**	.12	.19	.06	**.24**	.10	.14	.22	(.77)		
	Conscientiousness	.09	.09	**.30**	.14	.09	.08	**.20**	.06	.24	.30	(.86)	
	Emotional stability	.14	.16	.08	**.33**	.13	.12	.06	**.19**	.44	.38	.29	(.78)

paragraph (with correlations of .40 and .34). Thus, the correlation between measures that share trait variance but do not share method variance (the monotrait-heteromethod correlations) should be larger than correlations between measures that share neither trait variance nor method variance (the heterotrait-heteromethod correlations).

An even more stringent requirement for convergent and discriminant validity evidence is that the self-report measure of social skill should be more highly correlated with other measures of social skill than with self-report measures of other traits. The MTMMM in Table 9.3 shows that, as would be expected, the monotrait-heteromethod correlations are generally larger than the heterotrait-monomethod correlations, reflecting associations between measures of different constructs assessed through the same method (inside the solid-line triangles). The data in the MTMMM in Table 9.3 provide mixed evidence in terms of these associations. Although the correlations between the self-report measure of social skill and the self-report measures of impulsivity and conscientiousness are relatively low (.14 and .20, respectively), the correlation between the self-report measure of social skill and the self-report measures of emotional stability is relatively high, at .35. Thus, the self-report measure of social skill overlaps with a self-report measure of emotional stability to the same degree that it overlaps with other measures of social skill. This finding might raise concerns about the discriminant validity of the self-report measure that is supposed to assess social skill. Thus, the correlation between measures that share trait variance but do not share method variance (the monotrait-heteromethod correlations) should be larger than correlations between measures that do not share trait variance but do share method variance (the heterotrait-monomethod correlations). Ideally, the researchers would like to see even larger monotrait-heteromethod correlations than those in Table 9.3 and even smaller heterotrait-monomethod correlations.

In sum, an MTMMM analysis, as developed by Campbell and Fiske (1959), provides useful guidelines for evaluating construct validity. By carefully considering the important effects of trait variance and method variance on correlations among measures, researchers can use the logic of an MTMMM analysis to gauge convergent and discriminant validity. In the decades since Campbell and Fiske published their highly influential work, researchers interested in measurement have developed even more sophisticated ways of statistically analyzing data obtained from an MTMMM study. For example, Widaman (1985) developed a strategy for using factor analysis to analyze MTMMM data. Although such procedures are beyond the scope of our discussion, readers should be aware that psychometricians continue to build on the work by Campbell and Fiske.

Despite the strong logic and widespread awareness of the approach, the MTMMM approach to evaluating convergent and discriminant validity evidence does not seem to be used very frequently. For example, we conducted a quick review of the articles published in the 2005 volume of *Psychological Assessment,* which is a research journal published by the American Psychological Association (APA). The journal is intended to present "empirical research on measurement and evaluation relevant to the broad field of clinical psychology" (APA, n.d.). In our survey, we identified 13 articles claiming to present evidence related to convergent and discriminant validity, or construct validity more generally. Of these 13 articles, only 2 used an MTMMM approach. Furthermore, the 2 articles used multiple measurement occasions as the multiple "methods." That is, participants completed the same measure at different times, providing the same method of measurement at two or three different times. Although this review is admittedly limited and informal, it underscores our impressions of the (in)frequency with which MTMMM analyses are used.

Regardless of the frequency of its use, the MTMMM has been an important development in the understanding and analysis of convergent and discriminant validity evidence. It has shaped the way that many people think about construct validity, and it is an important component of a full understanding of psychometrics.

Quantifying Construct Validity

The final method that we will discuss for evaluating convergent and discriminant validity evidence is a more recent development. Westen and Rosenthal (2003) outlined a procedure that they called "quantifying construct validity" (QCV), in which researchers quantify the degree of "fit" between (a) their theoretical predictions for a set of convergent and discriminant correlations and (b) the set of correlations that are actually obtained.

At one level, this should sound familiar, if not redundant! Indeed, an overriding theme in our discussion of construct validity is that the theoretical basis of a construct guides the study and interpretation of validity evidence. For example, in the previous sections, we have discussed various ways that researchers identify the criterion variables used to evaluate convergent and discriminant validity evidence, and we have emphasized the importance of interpreting validity correlations in terms of conceptual relevance to the construct of interest.

However, evidence regarding convergent and discriminant validity often rests on rather subjective and impressionistic interpretations of validity correlations. In our earlier discussion of the "sets of correlations" approach to convergent and discriminant validity evidence, we stated that researchers often "eyeball" the correlations and make a somewhat subjective judgment about the degree to which the correlations match what would be expected on the basis of the nomological network surrounding the construct of interest. We also stated that researchers often judge the degree to which the pattern of convergent and discriminant correlations "makes sense" in terms of the theoretical basis of the construct being assessed by a test. But what if one researcher's judgment of what makes sense does not agree with another's judgment? And exactly how strongly do the convergent and discriminant correlations actually fit with the theoretical basis of the construct?

Similarly, when examining the MTMMM correlations, we asserted that some correlations were "generally larger" or "noticeably lower" than others. We must admit that we tried to sneak by without defining what we meant by "generally larger" and without discussing exactly *how much* lower a correlation should be in order to be considered "noticeably" lower than another. In sum, although the correlations themselves are precise estimates of association, the interpretation of the overall pattern of convergent and discriminant correlations often has been done in somewhat imprecise and subjective manners.

Given the common tendency to rely on somewhat imprecise and subjective evaluations of patterns of convergent and discriminant correlations, the QCV procedure was designed to provide a precise and objective quantitative estimate of the support provided by the overall pattern of evidence. Thus, the emphasis on precision and objectivity is an important difference from the previous strategies. The QCV procedure is intended to provide an answer to a single question in an examination of the validity of a measure's interpretation—"does this measure predict an array of other measures in a way predicted by theory?" (Westen & Rosenthal, 2003, p. 609).

There are two complementary kinds of results obtained in a QCV analysis. First, researchers obtain two effect sizes representing the *degree of fit* between the actual pattern of correlations and the predicted pattern of correlations. These effect sizes, called $r_{alerting\text{-}CV}$ and $r_{contrast\text{-}CV}$, are correlations themselves, ranging between –1 and +1. We will discuss the nature of these effect sizes in more detail, but for both, large positive effect sizes indicate that the actual pattern of convergent and discriminant correlations closely matches the pattern of correlations predicted on the basis of the conceptual meaning of the constructs being assessed. The second kind of result obtained in a QCV analysis is a test of statistical significance. The significance test indicates whether the degree of fit between actual and predicted correlations is likely to have occurred by chance. Researchers conducting a validity study using the QCV procedure will hope to obtain large values for the two effect sizes as well as statistically significant results.

The QCV procedure can be summarized in three phases. First, researchers must generate clear predictions about the pattern of convergent and discriminant validity correlations that they would expect to find. They must think carefully about the criterion measures included in the study, and they must form predictions for each one, in terms of its correlation with the primary measure of interest. For example, Furr and his colleagues (Furr, Reimer, & Bellis, 2004; Nave & Furr, 2006) developed a measure of social motivation, which was defined as a person's general desire to make positive impressions on other people. To evaluate the convergent and discriminant validity of the scale, participants were asked to complete the Social Motivation scale along with 12 additional personality questionnaires. To use the QCV procedure, Furr et al. (2004) needed to generate predictions about the correlations that would be obtained between the Social Motivation scale and the 12 additional scales. They did this by recruiting five professors of psychology to act as "expert judges." The judges read descriptions of each scale, and each one provided predictions about the correlations. The five sets of predictions were then averaged to generate a single set of predicted correlations.

The criterion scale labels and the predicted correlations are presented in Table 9.4. Thus, the conceptually guided predictions for convergent and discriminant correlations are stated concretely. For example, judges predicted that social motivation would be relatively strongly correlated with public self-consciousness (e.g., "I worry about what people think of me" and "I want to amount to something

Table 9.4 Example of the Quantifying Construct Validity Process

Criteria Scales	*Predicted Correlations*	*Actual Correlations*	*z-Transformed Correlations*
Dependence	.58	.46	.50
Machiavellianism	.24	.13	.13
Distrust	–.04	–.24	–.24
Resourcefulness	.06	–.03	–.03
Self-efficacy	–.04	.12	.12
Extraversion	.18	.03	.03
Agreeableness	.36	.39	.41
Complexity	.08	.06	.06
Public self-consciousness	.64	.51	.56
Self-monitoring	.56	.08	.08
Anxiety	.36	.24	.24
Need to belong	.56	.66	.79

special in others' eyes") and the need to belong (e.g., "I need to feel that there are people I can turn to in times of need" and "I want other people to accept me"). The judges expect that people who profess a desire to make positive impressions on others should report the tendency to worry about others' impressions and to need to be accepted by others. Conversely, the judges did not believe that social motivation scores would be associated with variables such as distrust and complexity, reflecting predictions of discriminant validity.

In the second phase of the QCV procedure, researchers collect data and compute the actual convergent and discriminant validity correlations. Of course, these correlations reflect the degree to which the primary measure of interest is *actually* associated with each of the criterion variables. For example, Furr et al. (2004) computed the correlations between the Social Motivation scale and each of the 12 criterion variables included in their study. As shown in Table 9.4, these correlations ranged from –.24 to .51. Participants who scored high on the Social Motivation scale tended to report relatively high levels of public self-consciousness and the need to belong. In addition, they tended to report relatively low levels of distrust, but they showed no tendency to report high or low levels of complexity or extraversion.

In the third phase, researchers quantify the degree to which the actual pattern of convergent and discriminant correlations fits the predicted pattern of correlations. A close fit provides good evidence of validity for the intended interpretation of the test being evaluated, but a weak fit would imply poor validity. As described earlier, the fit is quantified by two kinds of results—effect sizes and a significance test.

The two effect sizes reflect the amount of evidence of convergent and discriminant validity as a matter of degree. The $r_{alerting\text{-}CV}$ effect size is the correlation between the set of predicted correlations and the set of actual correlations. A large positive $r_{alerting\text{-}CV}$ would indicate that the correlations that judges predicted to be relatively large were indeed the ones that actually were relatively large, and it indicates that the correlations that judges predicted to be relatively small were indeed the ones that actually were relatively small. Take a moment to examine the correlations in Table 9.4. Note, for example, that the judges predicted that dependence, public self-consciousness, self-monitoring, and the need to belong would have the largest correlations with social motivation. In fact, three of these four scales did have the largest correlations. Similarly, the judges predicted that distrust, resourcefulness, self-efficacy, and complexity would have the smallest correlations with social motivation. In fact, three of these four scales did have the smallest correlations (relative to the others). Thus, the actual correlations generally matched the predictions made by the judges. Consequently, the $r_{alerting\text{-}CV}$ value for the data in Table 9.4 is .79, a large positive correlation. In actuality, the $r_{alerting\text{-}CV}$ value is computed as the correlation between the predicted set of correlations and the set of "*z*-transformed" actual correlations. The *z* transformation is done for technical reasons regarding the distribution of underlying correlation coefficients. For all practical purposes, though, the $r_{alerting\text{-}CV}$ effect size simply represents the degree to which the correlations that are predicted to be relatively high (or low) are the correlations that actually turn out to be relatively high (or low).

Although its computation is more complex, the $r_{contrast\text{-}CV}$ effect size is similar to the $r_{alerting\text{-}CV}$ effect size in that large positive values indicate greater evidence of convergent and discriminant validity. Specifically, the computation of $r_{contrast\text{-}CV}$ adjusts for the intercorrelations among the criterion variables and for the absolute level of correlations between the main test and the criterion variables. For the data collected by Furr et al. (2004), the $r_{contrast\text{-}CV}$ value was approximately .68, again indicating a high degree of convergent and discriminant validity. As the QCV procedure is a relatively recent development, there are no clear guidelines about how large the effect sizes should be in order to be interpreted as providing evidence of adequate validity. At this point, we can say simply that higher effect sizes offer greater evidence of validity.

In addition to the two effect sizes, the QCV procedure provides a test of statistical significance. Based on a number of factors, including the size of the sample and the amount of support for convergent and discriminant validity, a *z* test of significance indicates whether the results were likely to have been obtained by chance.

Although the QCV approach is a potentially useful approach to estimating convergent and discriminant evidence, it is not perfect. For example, low effect sizes (i.e., low values for $r_{alerting\text{-}CV}$ and $r_{contrast\text{-}CV}$) might not indicate poor evidence of validity. Low effect sizes could result from an inappropriate set of predicted correlations. If the predicted correlations are poor reflections of the nomological network surrounding a construct, then a good measure of the construct will produce actual correlations that do not match predictions. Similarly, a poor choice of criterion variables could result in low effect sizes. If few of the criterion variables used in the validity study are associated with the main test of interest, then they do not represent the nomological network well. Thus, the criterion variables selected for a QCV analysis should represent a range of strong and weak associations, reflecting a clear pattern of convergent and discriminant evidence. Indeed, Westen and Rosenthal (2005) point out that "one of the most important limitations of all fit indices is that they cannot address whether the choice of items, indicators, observers, and so forth was adequate to the task" (p. 410).

In addition, the QCV procedure has been criticized for resulting in "high correlations in cases where there is little agreement between predictions and observations" (G. T. Smith, 2005, p. 404). That is, researchers might obtain apparently large values for $r_{alerting\text{-}CV}$ and $r_{contrast\text{-}CV}$ even when the observed pattern of convergent and discriminant validity correlations does not match closely the actual pattern of convergent and discriminant validity correlations. Westen and Rosenthal (2005) acknowledge that this might be true in some cases; however, they suggest that the QCV procedures are "aids to understanding" and should be carefully scrutinized in the context of many conceptual, methodological, and statistical factors (p. 411).

We have outlined several strategies that can be useful in many areas of test evaluation, but there is no single perfect method or statistic for estimating the overall convergent and discriminant validity of test interpretations. Although it is not perfect, the QCV does offer several advantages over some other strategies. First, it forces researchers to consider carefully the pattern of convergent and discriminant associations that would make theoretical sense, on the basis of the construct in question. Second, it forces researchers to make explicit predictions about the pattern of associations. Third, it retains the focus on the measure of primary interest. Fourth, it provides a single interpretable value reflecting the overall degree to which the pattern of predicted associations matches the pattern of associations that is actually obtained, and it provides a test of statistical significance. Used with care, the QCV is an important addition to the toolbox of validation.

Factors Affecting a Validity Coefficient

The strategies outlined above are used to accumulate and interpret convergent and validity evidence. To some extent, all the strategies rest on the size of validity coefficients—statistical results that represent the degree of association between a test of interest and one or more criterion variables. In this section, we address some important factors that affect the validity coefficients that are obtained in the evaluation of convergent and discriminant validity.

We have emphasized the correlation as a coefficient of validity because of its interpretability as a standardized measure of association. Although other statistical values can be used to represent associations between tests and criterion variables (e.g., regression coefficients), most such values are built on correlation coefficients. Thus, our discussion centers on some of the key psychological, methodological, psychometric, and statistical factors affecting correlations between tests and criterion variables.

Associations Between Constructs

One factor affecting the correlation between measures of two constructs is the "true" association between those constructs. If two constructs are strongly associated with each other, then measures of those constructs will likely be highly correlated with each other. Conversely, if two constructs are unrelated to each other, then measures of those constructs will probably be weakly correlated with each other. Indeed, when we conduct research in general, we intend to interpret the observed correlations that we obtain (i.e., the correlations between the measured variables in our study) as approximations of the true correlations between the constructs in which we are interested. When we conduct validity research, we predict that two measures will be correlated because we believe that the two constructs are associated with each other.

Measurement Error and Reliability

In earlier chapters, you learned about the conceptual basis, the estimation, and the importance of reliability as an index of (the lack of) measurement error. As we discussed, one important implication of measurement error is its effect on correlations between variables—measurement error reduces, or attenuates, the correlation between measures. Therefore, measurement error affects validity coefficients.

As we saw in earlier chapters, the correlation between tests (say, X_1 and Y_2) of two constructs is a function of the true correlation between the two constructs and the reliabilities of the two tests:

$$r_{x_o y_o} = r_{x_t y_t} \sqrt{\mathrm{R}_{xx} \mathrm{R}_{yy}} \tag{9.1}$$

In this equation, $r_{x_o y_o}$ is the correlation between the two tests. It is the validity correlation between the primary test of interest and the test of a criterion variable. In addition, $r_{x_t y_t}$ is the true correlation between the two constructs, R_{xx} is the reliability of the test of interest, and R_{yy} is the reliability of the test of the criterion variable.

For example, in their examination of the convergent validity evidence for their measure of social motivation, Furr et al. (2004; Nave & Furr, 2006) were interested in the correlation between social motivation and public self-consciousness. Imagine that the true correlation between the constructs is .60. What would the actual validity correlation be if the two tests had poor reliability? If the social motivation test had a reliability of .63 and the public self-consciousness test had a reliability of .58, then the actual validity coefficient obtained would be only .36:

$$r_{x_o y_o} = .60\sqrt{.63}\sqrt{.58},$$

$$r_{x_o y_o} = .60(.604),$$

$$r_{x_o y_o} = .36.$$

Recall that, to evaluate convergent validity, researchers should compare their correlations with the correlations that they would expect based on the constructs being measured. In this case, if Furr et al. (2004) were expecting to find a correlation close to .60, then they might be relatively disappointed with a validity coefficient of "only" .36. Therefore, they might conclude that their test has poor validity as a measure of social motivation.

Note that the validity coefficient is affected by two reliabilities: (a) the reliability of the test of interest and (b) the reliability of the criterion test. Thus, the primary test of interest could be a good measure of the intended construct, but the validity coefficient could appear to be poor. If the social motivation test had a reliability of .84 but the public self-consciousness test had a very poor reliability of .40, then the actual validity coefficient obtained would be only .35:

$$\begin{aligned} r_{x_o y_o} &= .60\sqrt{.84}\sqrt{.40}, \\ r_{x_o y_o} &= .60(.580), \\ r_{x_o y_o} &= .35. \end{aligned}$$

Therefore, when evaluating the size of a validity correlation, it is important to consider both the reliability of the primary test of interest and the reliability of the criterion test. If either one or both is relatively weak, then the resulting validity correlation is likely to appear relatively weak. This might be a particularly subtle consideration for the criterion variable. Even if the primary test of interest is a good measure of its intended construct, we might find poor validity correlations. That is, if the criterion measures that we use are poor, then we are unlikely to find good evidence of validity even if our test of interest is actually a good measure of its intended construct! This important issue is easy to forget.

Restricted Range

Recall that a correlation coefficient reflects covariability between two distributions of scores. That is, it represents the degree to which variability in one distribution of scores (e.g., scores on a test to be validated) corresponds with variability in another distribution of scores (e.g., scores on a test of a criterion variable). The amount of variability in one or both distributions of scores can affect the correlation between the two sets of scores. Specifically, a correlation between two variables can be reduced if the range of scores in one or both variables is artificially limited or restricted.

A classic example of this is the association between SAT scores and academic performance. Earlier, we discussed the fact that much evidence for the quality of the SAT scores rests on the correlation between SAT scores and academic performance, as measured by college grade point average (GPA). The marketers of the SAT would like to demonstrate that people who score relatively high on the SAT tend to have relatively good performance in college. Implicitly, this demonstration requires that people who score relatively low on the SAT tend to have relatively poor performance in college. To demonstrate this kind of association, researchers would need to demonstrate that variability in the distribution of SAT scores corresponds with variability in the distribution of college GPAs. However, the ability to demonstrate this association is minimized by restricted range in two ways.

First, range restriction exists in GPA as a measure of academic performance. In most colleges, GPA can range only between 0.0 and 4.0. The worst that any student can do is a GPA of 0.0, and the best that any student can do is 4.0. But does this 4-point range in GPA really reflect the full range of possible academic performance? Consider two students who both do well in classes and earn A's in all of their courses. Although Leo did perform well, he barely earned an A in each of his courses. So, he "squeaked by" with a 4.0, and the 4.0 in a sense represents the upper limit of his academic performance. Mary also performs well, earning A's in all of her courses. But Mary outperformed every other student in each of her courses. In each course, she was the only one to earn an A on any tests, and she had clearly mastered all the material on each and every assignment that her professors graded. So, Mary also received a 4.0, but her 4.0 in a sense underestimates her academic ability. She had mastered all the material so well that her professors wish that they could give her grades higher than an A. Although Leo and Mary received the same "score" on the measure of academic performance (i.e., they both have 4.0 GPAs), they actually differ in the quality of their performance. Leo fully earned his 4.0 and should be proud of it, but the professors would probably agree that Mary outperformed him. Thus, the 4-point GPA scale restricts the range of measurement of academic performance.

Note that GPA is restricted in both directions—on the high end and on the low end. Consider Jenny and Bruce. Although both Jenny and Bruce failed all of their classes, Bruce nearly passed each class. On the other hand, Jenny wasn't even close to passing any classes. So, both Bruce and Jenny earned a GPA of 0.0, but in a sense, Bruce had greater academic performance than Jenny. In terms of test grades, homework grades, and paper grades, Bruce outperformed Jenny (i.e., he received 59s on each assignment, while she received scores on the 30s on each assignment). Despite the difference in their performance during the semester, the GPA scale "bottoms out" at 0.0, so Jenny cannot receive a lower GPA than Bruce.

The scatterplot in Figure 9.1 shows a hypothetical data set for 5,000 students. This scatterplot presents the idealized association between SAT scores and "unrestricted" college GPA. That is, it presents scores for students whose academic performance is not restricted by a 4-point GPA scale. Notice that some unrestricted GPA scores fall below 0.0 on the plot, reflecting differences between students like Jenny and Bruce. Notice also that some GPA scores fall above 4.0, reflecting differences between students like Leo and Mary. For the data displayed in Figure 9.1, the correlation between SAT and GPA was .61. This indicates that students who received relatively low SAT scores tended to have relatively low "unrestricted" GPAs.

But of course, GPA actually is restricted. Therefore, students whose academic performance might merit a 5.0 or a 6.0 can earn only a 4.0. Similarly, students whose academic performance might merit a GPA below zero cannot actually

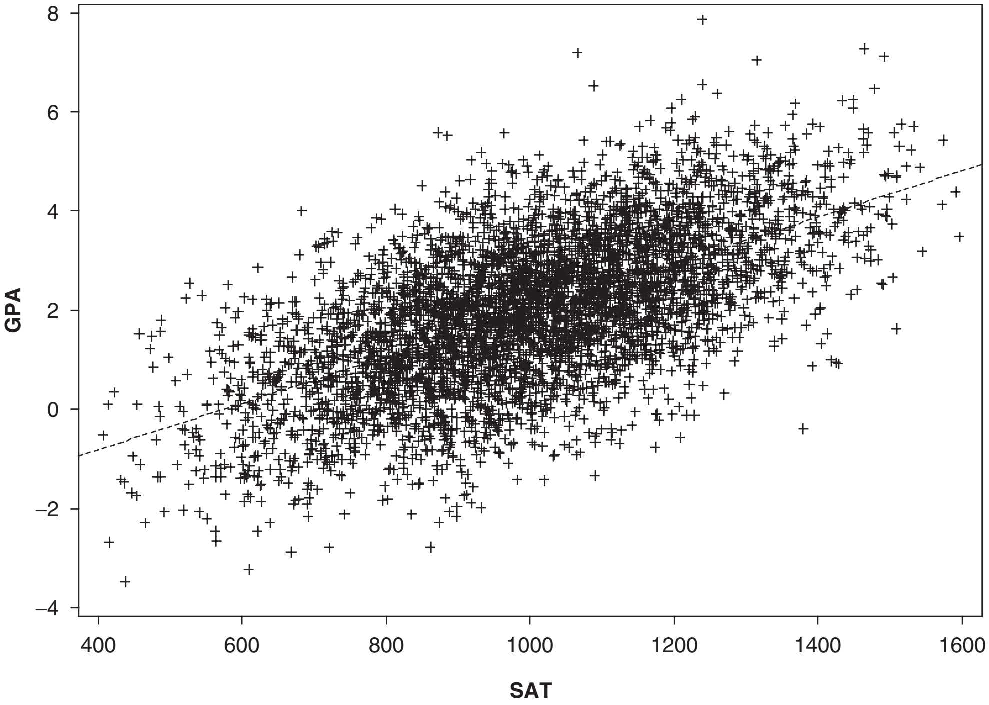

Figure 9.1 Scatterplot of SAT and "Unrestricted" College GPA

receive less than a zero. So, all those students who might, in an abstract sense, deserve GPAs above 4.0 (or below 0) will in reality receive a GPA of 4.0 (or 0).

The scatterplot in Figure 9.2 shows the data for the same 5,000 students, based on the "restricted" GPA scores. So, note that there are no GPA scores above 4.0—scores are "maxed out" at 4.0. And note that there are no GPA scores below 0.0—scores are bottomed out at 0. This scatterplot appears to be more compressed, and the association between SAT and GPA is not as clear as it was in the first scatterplot. Consequently, for the data displayed in Figure 9.4, the correlation between SAT and GPA was reduced, a bit, to .60. Thus, the restriction of range in GPA scores has a slightly diminishing effect on the correlation.

A second way in which range restriction minimizes the ability to demonstrate the association between SAT scores and academic performance is in the number of people who actually obtain college GPAs. That is, students with very low SAT scores are much less likely to be admitted to college than are students with higher SAT scores. If we were to conduct a real study of the association between SAT scores and academic performance, we would probably be limited to a subsample of all the students who have SAT scores. This is because we would be limited to only those students who took the SAT and who were admitted to college. For better or for worse, not all students who take the SAT are admitted to college. In our hypothetical data set, nearly 400 "students" had SAT scores below 700. In reality, these students might not be admitted to college; therefore, they would never actually have a college GPA score.

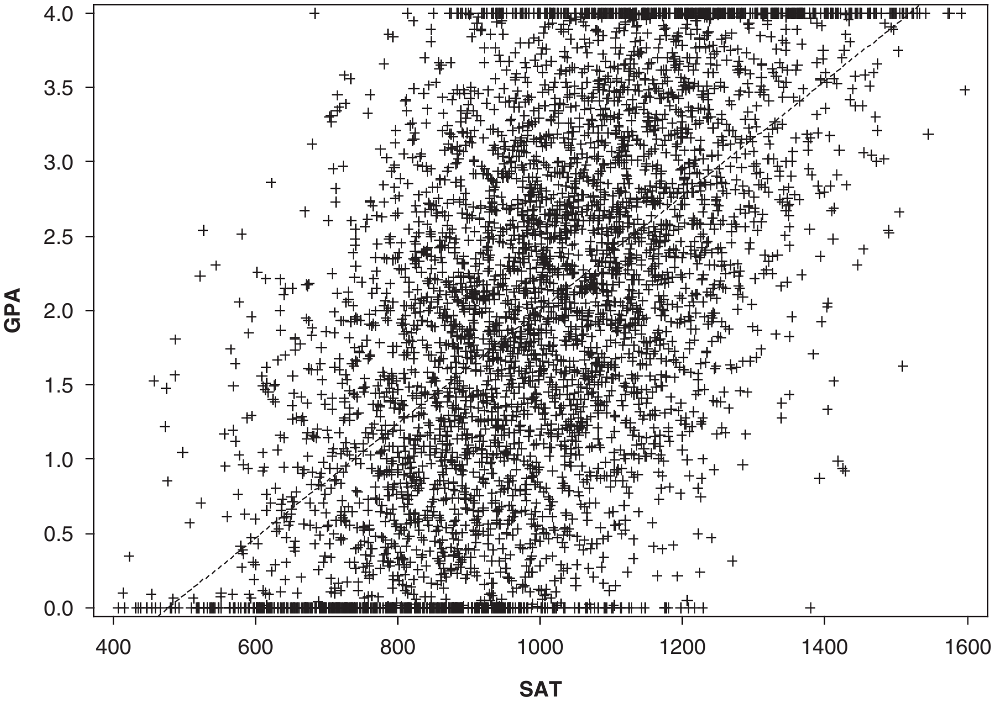

Figure 9.2 Scatterplot of SAT and Restricted College GPA

The scatterplot in Figure 9.3 shows the data for the remaining 4,600 students, with SAT scores greater than 700. Note that there are no people with SAT scores below 700. We are assuming that most if not all of those people would not be admitted to college and thus could not be included in an analysis of the association between SAT scores and college GPAs. Again, this scatterplot appears to be more compressed than the previous two. Consequently, for the data displayed in Figure 9.3, the correlation between SAT and GPA was reduced even more, to .55.

In sum, the SAT and GPA example illustrates range restriction and its effect on validity correlations. When evaluating the quality of a psychological measure, we often depend on correlations (or other statistical values that are based on correlations) to reflect the degree of convergent and discriminant validity. When searching for convergent evidence, we expect to find strong correlations. However, we need to be aware that restricted range can reduce the correlations that are actually obtained in a validity study. In the current example, the correlation between SAT and GPA was affected by restriction in two ways, and it was somewhat smaller than an "unrestricted" correlation between SAT scores and academic performance. Although the effect of range restriction might not be dramatic in the current example, being aware of the phenomenon improves our ability to interpret validity coefficients.

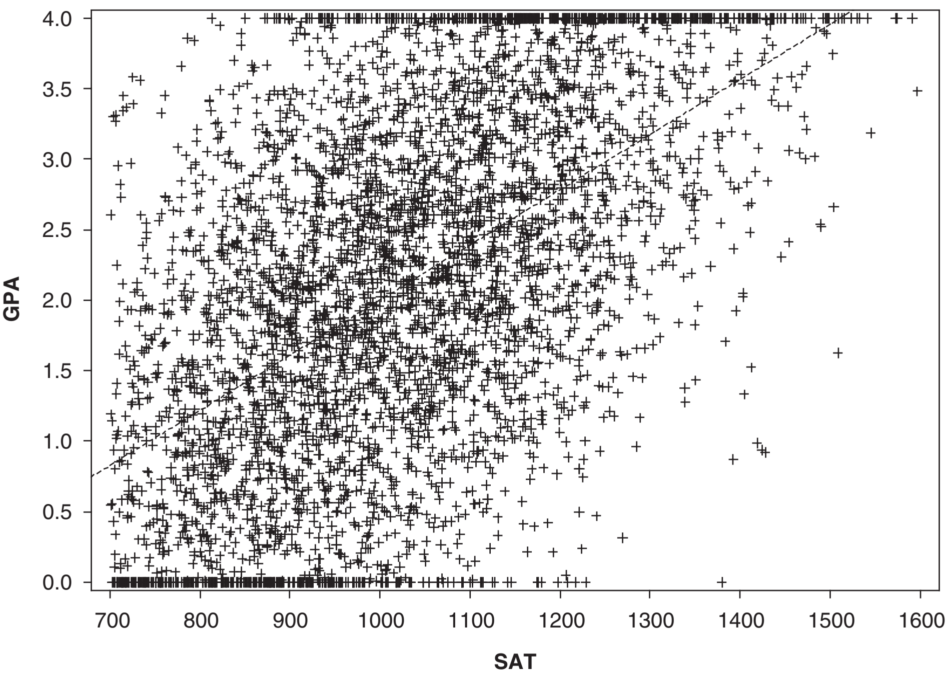

Figure 9.3 Scatterplot of Restricted SAT and Restricted College GPA

Relative Proportions

Imagine that we developed a self-report inventory to measure depression. And imagine that we would like to evaluate its convergent quality by correlating its scores with diagnoses made by trained clinical psychologists. We recruit a sample of participants who complete our inventory and who are interviewed by clinicians. The clinicians provide a diagnosis for each participant, labeling each participant as either depressed or nondepressed. Thus, our main test of interest is on a continuous scale, and the criterion variable (i.e., diagnosis) is a dichotomous categorical variable. We would hope to find that participants' scores on our new inventory are strongly correlated with clinicians' ratings.

The size of the validity correlation between inventory scores and clinicians' diagnoses is influenced by the proportion of participants who are diagnosed as having depression. Let us examine this influence. If we were computing this correlation, each participant would have scores on two variables—depression inventory score and diagnostic category. See Table 9.5 for hypothetical data illustrating this example. Obviously, the depression inventory scores are already on a quantitative scale (let us say that scores can range from 0 to 30). However, the diagnostic category variable must be quantified so that we can compute the validity. To do this, we assign one value to all participants diagnosed as nondepressed and another value to all participants diagnosed as depressed. These values could be 1 and 2, or 1 and 10, or –1000 and +1000, or any other pair of numbers (as long as all the people in each group receive the same value). For our purposes, we will code the "nondepressed" group as 1 and the depressed group as 2 (see Table 9.5).

Table 9.5 Data Illustrating the Effect of Relative Proportions on Validity Coefficients

Participant	*Depression Inventory Score*	*Diagnosis*	*Diagnosis Code*
1	6	Nondepressed	1
2	5	Nondepressed	1
3	7	Nondepressed	1
4	1	Nondepressed	1
5	11	Nondepressed	1
6	9	Nondepressed	1
7	3	Nondepressed	1
8	6	Nondepressed	1
9	4	Nondepressed	1
10	8	Nondepressed	1
11	10	Nondepressed	1
12	2	Nondepressed	1
13	5	Nondepressed	1
14	7	Nondepressed	1
15	6	Nondepressed	1
16	10	Depressed	2
17	15	Depressed	2
18	5	Depressed	2
19	8	Depressed	2
20	12	Depressed	2
Mean	7.00		.25
Standard Deviation	3.39		.43
Covariance		.75	
Correlation		.51	

Recall from a previous chapter that the correlation between two variables is the covariance between the two variables divided by the product of their two standard deviations. For a correlation between one continuous variable and one dichotomous variable (r_{CD}),

$$r_{CD} = \frac{c_{CD}}{s_C s_D}, \tag{9.2}$$

where c_{CD} is the covariance between the two variables, s_C is the standard deviation of the continuous variable, and s_D is the standard deviation of the dichotomous variable.

Two of these terms are directly affected by the proportion of observations in the two groups defined by the dichotomous variable. Assuming that the groups are coded 1 (for Group 1) and 2 (for Group 2), then the covariance is

$$c_{CD} = p_1 p_2(\bar{C}_2 - \bar{C}_1), \tag{9.3}$$

where p_1 is the proportion of participants who are in Group 1, p_2 is the proportion of participants who are Group 2, $\bar{C}_1$ is the mean of the continuous variable for the participants in Group 1, and $\bar{C}_2$ is the mean of the continuous variable for the participants in Group 2. In our data set, 15 of the 20 participants are in the nondepressed diagnostic group (Group 1), and 5 are in the depressed group (Group 2). Thus, the two proportions are .75 (15/20 = .75) and .25 (5/20 = .25). In addition, the average score on the depression inventory is 6 for the nondepressed group, and it is 10 for the depressed group. Thus, the covariance is

$$\begin{aligned} c_{CD} &= (.75)\ (.25)\ (10 - 6), \\ c_{CD} &= (.1875)\ (4), \\ c_{CD} &= .75. \end{aligned}$$

The standard deviation of the dichotomous variable is the second term affected by the proportion of observations in the two groups defined by the dichotomous variable. Again, assuming that the groups are coded 1 (for Group 1) and 2 (for Group 2), then this term is

$$s_D = \sqrt{p_1 p_2}. \tag{9.4}$$

For the data in Table 9.5, the standard deviation of the dichotomous "diagnosis" variable is

$$\begin{aligned} s_D &= \sqrt{(.75)(.25)}, \\ s_D &= .433. \end{aligned}$$

Taking these terms into account, the equation for the correlation can be reframed and simplified to show the direct influence of the relative proportions:

$$r_{CD} = \frac{p_1 p_2(\overline{C}_2 - \overline{C}_1)}{s_C\sqrt{p_1 p_2}},$$
$$r_{CD} = \frac{\sqrt{p_1 p_2}(\overline{C}_2 - \overline{C}_1)}{s_C}. \tag{9.5}$$

For the example data in Table 9.5, the validity correlation is

$$r_{CD} = \frac{\sqrt{(.75)(.25)}(10-6)}{3.39},$$
$$r_{CD} = \frac{1.72}{3.39},$$
$$r_{CD} = .51.$$

This correlation is positive and fairly strong, indicating that those participants with relatively "high scores" on the diagnosis variable tended to obtain higher scores on the depression inventory than did those participants with relatively "low scores" on the diagnosis variable. Recall that we coded the diagnosis variable so that participants who were diagnosed as depressed had higher "diagnosis" scores (i.e., a score of 2) than participants who were diagnosed as nondepressed. Therefore, we can interpret the correlation as showing that participants diagnosed as depressed (i.e., those with relatively high scores on the diagnosis variable) tended to obtain higher scores on the depression inventory than did the participants diagnosed as nondepressed (i.e., those with relatively low scores on the diagnosis variable).

Equation 9.5 reveals the influence of group proportions on validity correlations. All else being equal, equally sized groups will allow larger correlations than will unequal groups. If two groups are equally sized, then the two proportions are .5 and .5. The product of these two proportions (.5 × .5 = .25) is the maximum of any two proportions. That is, any other pair of proportions will produce a product less than .25, and the greater the disparity in group sizes, the lower the product (e.g., .40 × .60 = .24, .10 × .90 = .09). And all else being equal, lower products will produce lower correlations.

In sum, a subtle factor that might affect some validity coefficients is the relative proportion of people in two groups. If a validity coefficient is based on the correlation between a continuous variable and a dichotomous (group) variable, then any disparity in the group sizes can reduce the resulting validity coefficient. This issue should be kept in mind when interpreting validity coefficients.

Method Variance

We discussed method variance in our earlier presentation of the multitrait-multimethod matrix (MTMMM). We will not say much more about it here;

however, method variance is an important consideration beyond its role in an MTMMM analysis. Any time that a researcher correlates test scores with scores from a different method of assessment, method variance is likely to reduce the correlation. Or perhaps more precisely stated, correlations between two different methods of assessment are likely to be smaller than correlations between measures from a single method of assessment.

This issue has an important implication for validity coefficients. When evaluating validity coefficients, we are more impressed with evidence from correlations between different methods of assessment than with evidence from a single method of assessment. For example, if we were evaluating a new self-report measure of social skill, we might correlate scores on the new measure with scores on self-report measures of extraversion. We might be happy to find a correlation of .40 between the measures, and we might interpret these results as evidence of convergent validity. After all, these results suggest that people who report having high social skill (based on our new measure) also report being relatively extraverted. Despite our satisfaction at finding these results, we would probably be even more enthusiastic if we had found a correlation of .40 between our self-report measure of social skill and an *acquaintance report* measure of extraversion. The result would be more compelling if we could say that people with high scores on our new measure of social skill are described as extraverted by their acquaintances.

Validity studies based solely on self-report are informative and common, but they are not perfect. Again, self-report data are relatively easy, inexpensive, and generally quite good, so we do not intend to imply that self-report data are inferior to data derived from other forms of measurement. However, correlations based solely on self-report questionnaires are potentially inflated due to shared method variance. In contrast, correlations that are based on data from different assessment methods are less likely to be artificially inflated. Thus, they provide an important complement to the more common reliance on self-report data. When interpreting correlations based on different methods, it is important to realize that they are likely to be smaller than correlations based solely on self-report data due to method variance.

Time

We have seen that construct validity is sometimes evaluated by examining the correlation between a test given at one point in time (e.g., SAT) and a criterion variable measured at a later point in time (e.g., college GPA). All else being equal, validity coefficients based on correlations between variables measured at different points in time (i.e., predictive validity correlations) are likely to be smaller than coefficients based on correlations between variables measured at a single point in time (i.e., concurrent validity correlations). Furthermore, it is likely that longer periods between two points in time will produce smaller predictive validity correlations.

Predictions of Single Events

An important factor that can affect validity coefficients is whether the criterion variable is based on an observation of a single event or on some kind of aggregation or accumulation of events. For example, imagine that you developed a questionnaire that you intend to interpret as a measure of extraversion. And imagine that you wished to gather convergent validity evidence by correlating its scores with observations of "talkativeness" in social interaction. Your understanding of the extraversion construct suggests that extraverted people should be relatively talkative in a social interaction, so you expect to find a moderate to large positive correlation between scores on your questionnaire and observations of talkativeness.

Let us say that you recruited a sample of 50 participants who completed your questionnaire and who engaged in a 5-minute social interaction with a stranger "partner" of the other sex. The partners then rated the participants on talkativeness, using a 1 to 10 scale, with high scores indicating greater talkativeness. You compute the correlation between your questionnaire and the talkativeness ratings, and you find only a small positive correlation. You are disappointed, and you feel compelled to conclude that your questionnaire is a poor measure of extraversion.

Before you decide to revise your measure or discard it entirely, you should consider the nature of your criterion variable. Specifically, you should remember that it was based on an observation of a single behavior (i.e., talkativeness) in a single social situation (i.e., a 5-minute interaction with another-sex stranger). Even beyond the issue of method variance, you should consider that there are many factors that could influence an individual's talkativeness in any one moment. What kind of mood was the individual in? How was the partner acting? Was there a task or a topic of conversation that inhibited the individual's talkativeness?

Chances are that your validity correlation could have been larger if you had gathered observations of your participants from several different interactions or over a longer period of time. For a variety of reasons, single events are less predictable than are aggregations of events or accumulations of observations (Epstein, 1979).

A particularly compelling example of the difficulty of predicting single events was provided by Abelson (1985). Some baseball players are paid tens of millions of dollars, partially because they have batting averages that are much higher than the average player. Presumably, owners and managers of baseball teams believe that players with high batting averages will be much more successful than players with low batting averages. That is, in any single at bat, the player with a high batting average should have a much greater chance of hitting the ball than a player with a low batting average. But is this actually true? How much variability in at-bat success is actually explained by batting average? Abelson examined baseball statistics to evaluate the association between batting average (scored from 0 to 1.0) and the chances of success at any single at bat.

Abelson's (1985, p. 132) analysis revealed what he interpreted as a "pitifully small" association between batting skill (as reflected in batting average) and success at a single at bat. In light of such a small statistical association, he considered why he, other statistical experts, other baseball fans, and even baseball managers believed that batting average is such an important issue. He concludes that "the individual batter's success is appropriately measured over a long season, not by the individual at bat" (p. 132). That is, although the ability to predict a single event (i.e., an individual at bat) is perhaps meager, what matters are the cumulative effects of many such events.

Even a meager level of predictability for any single event can produce a much more substantial level of predictability as those events accumulate.

In sum, single events—whether they are baseball at bats or a specific social behavior in a specific social situation—might be inherently unpredictable. In terms of validity coefficients, one must consider this issue in relation to the criterion variable. Is the criterion to be predicted a single event, such as a single observation of social behavior? Or is the criterion a cumulative variable, such as the average level of social behavior across many observations? Large validity coefficients are more likely to be obtained when the criterion variable is based on the accumulation or aggregation of several events than when it is based on only a single event.

Interpreting a Validity Coefficient

After a validity coefficient is obtained, it must be interpreted. Test developers, evaluators, and users must decide if validity coefficients are large enough to provide compelling evidence of convergence or if they are small enough to provide assurance of discriminant validity. Although it is a precise way of quantifying the degree of association between two measures, the correlation coefficient might not be highly intuitive. Particularly for newcomers to a field of study, the knowledge that a correlation is, for example, .40 is not always very informative. In our experience, the tendency seems to be for people to note that .40 seems far from a perfect correlation of 1.0, and thus they interpret it as quite small. Anything less than perfect is seen as a somewhat weak association.

This tendency could be problematic when evaluating a validity coefficient, particularly when discussing validity with someone who is not experienced with interpreting correlations. For example, the human resources director for a company might need to convince employers, test takers, or lawyers that a particular test is a valid predictor of job performance. To make her case, she cites research evidence showing a .40 correlation between test scores and job performance. As we know, this suggests that people who score relatively high on the test tend to exhibit relatively high job performance. However, her audience of employers, test takers, or lawyers might interpret this evidence quite differently. In fact, they might argue that a correlation of .40 is far from perfect, and they might even interpret it as evidence of the invalidity of the test! How could the human resources director convince others that the test is actually a useful and valid predictor?

As discussed above, issues such as the true correlation between constructs, method variance, relative proportions, and reliability are some key factors affecting the size of a validity coefficient. Several additional important issues become relevant in the overall interpretation of the size and meaning of a validity coefficient.

Squared Correlations and "Variance Explained"

In psychological research, a common practice is to interpret a squared correlation. A squared correlation between two variables is often interpreted as the proportion of variance in one variable that is explained or "accounted for" by the other. For example, if we found a correlation of .30 between social skill and self-esteem, we might interpret this as showing that 9% of the variance in self-esteem is explained by social skill (.30 squared is .09). Actually, we could also interpret this result as showing that 9% of the variance in social skill is explained by self-esteem.

The "variance explained" interpretation is appealing, given our earlier assertion that research in general (and psychometrics in particular) is concerned with measuring and understanding variability. The more variability in a phenomenon that we can explain or account for, the more we feel like we understand the phenomenon. Furthermore, the variance explained interpretation fits various statistical procedures such as regression and analysis of variance (ANOVA), which rely on partitioning or predicting variability. Thus, you will frequently read or hear researchers interpreting associations in terms of squared correlations and the amount of variance explained.

Despite the appeal of this approach, the "squared correlation" approach to interpreting associations has been criticized for at least three reasons. First, it is technically incorrect in some cases. Although the statistical basis of this argument is beyond the scope of our current discussion, Ozer (1985) argues that in some cases, the correlation itself and not the squared correlation is interpretable as the proportion of variation explained. Second, some experts point out that variance itself is on a nonintuitive metric. Recall from an earlier chapter that, as a measure of differences among a set of scores, variance is based on *squared* deviations from the mean. The variance has some nice statistical properties, but how are we to interpret squared differences from a mean? D'Andrade and Dart (1990) point out that thinking in terms of squared differences or distance is not usually meaningful—do you provide directions to your house by telling friends that it is 9 squared miles from the interstate? The squared correlation approach might be seen as a nonintuitive and therefore nonuseful perspective on the association between variables.

The third criticism of the squared correlation approach is the least technical but perhaps the most powerful of the three. Simply put, squaring a correlation makes the association between two variables seem too small. It is not uncommon to hear researchers bemoaning the fact that they have explained "only" 9% or 12% of the variance in a phenomenon. Or you might read criticism of a research finding that explains "only" 16% of the variance. Indeed, 9%, 12%, and 16% do not sound like great amounts of anything. After all, this implies that nearly 90% of the variance is unexplained, and that sounds like a lot! However, as we will discuss in a later section, 9%, 12%, or 16% of the variance in a phenomenon might be a meaningful and important amount of variance. This is particularly true if we are talking about the association between only two variables. For example, if we can use a single variable, such as social skill, to explain nearly 10% of the variability in an important and complex phenomenon such as self-esteem, then perhaps that is a pretty important association.

The baseball example provided by Abelson (1985) is also relevant here. Recall that Abelson's examination led him to conclude that the association between batting average and the chances of success at any single at bat was very small. In fact, his conclusion was based on analyses revealing that only *one third of 1%* of the variance in any single batting performance was explained by batting skill (as reflected in batting average). As discussed earlier, Abelson pointed out that the cumulative effect of many at bats could account for the general belief that batting average was an important indicator of batting skill. D'Andrade and Dart (1990) offer a different perspective in explaining the discrepancy between Abelson's effect size (an apparently very small percentage of variance) and the common wisdom (batting average is an important statistic). They suggest that the discrepancy partially results from the fact that percentage of variance is a poor measure of association. Commenting on a table provided by Abelson, they point out that his results could be legitimately interpreted as showing that the difference between a .220 batter and a .320 batter produces a 10% difference in the likelihood of getting a hit in any single at bat. D'Andrade and Dart acknowledge that "10% may not be huge," but they suggest that "those who bet to win like 10% edges. So do baseball managers" (p. 58).

The "squared correlation" or "variance explained" interpretation of validity coefficients is a common but potentially misleading approach. Although it fits the view of research and measurement as tied to variability, it has several technical and logical problems. Perhaps most critically, a "variance explained" approach tends to cast associations in a way that seems to minimize their size and importance.

Estimating Practical Effects: Binomial Effect Size Display, Taylor-Russell Tables, Utility Analysis, and Sensitivity/Specificity

One way of interpreting a correlation is by estimating its impact on "real-life" decision making and predictions. The larger a correlation is between a test and a criterion variable, the more successful we will be in using the test to make decisions about the criterion variable. Returning to our example of the human resources director, she might frame the question in terms of the success in using the test to make hiring decisions, in terms of predictions about job performance. That is, for people who score relatively high on the test, how often will she be correct in predicting that those people will exhibit relatively good job performance? How often will she be incorrect? There are at least four procedures that have been developed to present the implications of a correlation in terms of our ability to use the correlation to make successful predictions—the binomial effect size display (BESD; Rosenthal & Rubin, 1982), the Taylor-Russell tables (Taylor & Russell, 1939), utility analysis (Brogden & Taylor, 1950), and an analysis of test sensitivity and specificity (Loong, 2003).

The BESD is designed to illustrate the practical consequences of using a correlation to make decisions. Specifically, it is usually formatted to make predictions or decisions for a group of 200 people—100 of whom have relatively high scores on the test and 100 who have relatively low scores. How many of the high scorers are likely to perform well on a criterion variable? How many of the low scorers are likely to perform well? See Table 9.6a for a 2 × 2 table that reflects this issue. We can use the BESD procedure to show how many successful and unsuccessful predictions will be made on the basis of a correlation.

Let us start with a worst-case scenario of zero correlation between test and criterion. If test scores are uncorrelated with job performance, then we would have only a 50/50 success rate (see Table 9.6a). Among 100 people with relatively low scores on the test, 50 would perform relatively poorly in their jobs, and 50 would perform relatively well. Similarly, among the 100 people with relatively high scores on the test, 50 would perform relatively poorly in their jobs, and 50 would perform relatively well. When a test is uncorrelated with a criterion variable, using the test to make predictions is no better than flipping a coin. Certainly, the human resources director would reject a test that had a validity coefficient that produced a success rate no better than flipping a coin.

What about a scenario in which there is a nonzero correlation between test and criterion? If test scores are correlated with job performance, then we would be more successful than 50/50. Rosenthal and Rubin (1982) provide a way of illustrating exactly how much more successful we would be. Note that the 2 × 2 table presented in Table 9.6b is formatted so that Cell A corresponds to the number of people who have low test scores and who are predicted to perform poorly at work. To determine this value, we use the following formula:

$$\text{Cell A} = 50 + 100\ ({}^{r}\!/_{2}),$$

Table 9.6 Example of the Binomial Effect Size Display (BESD)

(a) For a correlation of $r = .00$

	Job Performance	
Test Score	*Poor*	*Good*
Low	50	50
High	50	50

(b) For a correlation of $r = .40$

	Job Performance	
Test Score	*Poor*	*Good*
Low	A 70	B 30
High	C 30	D 70

where *r* is the correlation between test and criterion. If test scores are correlated with job performance at $r = .40$, then we would predict that 70 people with low test scores would have poor job performance:

$$\begin{aligned}\text{Cell A} &= 50 + 100\ ({}^{.40}\!/_{2}),\\ \text{Cell A} &= 50 + 20,\\ \text{Cell A} &= 70.\end{aligned}$$

Our prediction for Cell B (the number of people with low test scores who are predicted to perform well) is

$$\begin{aligned}\text{Cell B} &= 50 - 100\ ({}^{r}\!/_{2}),\\ \text{Cell B} &= 50 - 100\ ({}^{.40}\!/_{2}),\\ \text{Cell B} &= 50 - 20,\\ \text{Cell B} &= 30.\end{aligned}$$

The predicted success rates for Cells C and D parallel those for Cells A and B:

$$\begin{aligned}\text{Cell C} = \text{Cell B} &= 50 - 100\ ({}^{r}\!/_{2}) = 30,\\ \text{Cell D} = \text{Cell A} &= 50 + 100\ ({}^{r}\!/_{2}) = 70.\end{aligned}$$

Now, based on the data presented in the BESD, let us consider the importance or utility of a correlation that is "only" .40. If the human resources director hired only applicants with relatively high test scores, then 70% of those applicants will turn out to exhibit good job performance, and only 30% will turn out to exhibit poor performance. A 70% success rate is not perfect, but it seems quite acceptable for complex phenomena such as work performance. Depending on the cost of training employees, employers might view a 70% success rate as very good indeed.

In sum, the BESD can be used to translate a validity correlation into a framework that is relatively intuitive. By framing the association as the rate of successful predictions, the BESD presents the association between a test and criterion in terms that most people are familiar with and can understand easily.

Despite the intuitive appeal of the BESD, it has been criticized as an estimate of the practical effects of a correlation (Hsu, 2004). One of the key criticisms of the BESD is that it automatically frames the illustration in terms of an "equal proportions" situation. That is, it is cast for a situation in which the number of people with low test scores is equal number to the number of people with high test scores. In addition, it is cast for a situation in which half the sample is "successful" on the criterion variable and half is unsuccessful. As described earlier in this chapter, the relative proportion of scores on a variable can affect the size of a correlation. Although the assumption of relative proportions might be reasonable in some cases, it might not be representative of many real-life situations. For example, the human resources director might hire only 10% of the sample, not 50%. In addition, strong job performance might be rather difficult to achieve, perhaps only a 20% chance.

For situations in which the equal proportions assumption is untenable, we can examine tables prepared by Taylor and Russell (1939). These tables were designed to inform selection decisions, and they provide the probability that a selection decision based on an "acceptable" test score will result in successful performance on the criterion. As with the BESD, the Taylor-Russell tables cast the predictor (test) and outcome scores as dichotomous variables. For example, the human resources director will conceive of test scores as either passing or failing, in terms of a hiring decision. In addition, she will conceive of the job performance criterion as either successful performance or unsuccessful performance. The key difference between the BESD and the Taylor-Russell tables is that the Taylor-Russell tables can accommodate decisions that are based on various proportions both for passing/failing on the test and for successful/unsuccessful performance.

To use the Taylor-Russell tables, we need to identify several pieces of information. First, what is the size of the validity coefficient? Second, what is the selection proportion—the proportion of people who are going to be hired? That is, are 10% of applicants going to be hired (leaving 90% not hired), or will 30% be hired? Third, what is the proportion of people who would have "successful" criterion scores, if the selection was made without the test? That is, assuming that hires were made without regard to the test scores, how many employees would achieve successful job performance?

With these three pieces of information, we can check the Taylor-Russell tables to estimate the proportion of people with acceptable scores who go on to have successful performance. For example, if we knew that 10% of a sample would be hired (a selection proportion of .10) and that the general rate of successful performance was 60% (a success proportion of .60), then we could estimate the benefit of using a test to make the selection decisions. If the applicant screening test has a validity coefficient of .30, then the Taylor and Russell tables tell the human resources director that 79% of the applicants selected on the basis of the test would show successful job performance. Note that this percentage is greater than the general success rate of 60%, which is the success rate that is estimated to occur if hires were made without the use of test scores. So, the human resources director concludes that the test improves successful hiring by 19%.

The Taylor-Russell tables have been popular in industrial/organizational psychology, in terms of hiring decisions. Our goal in describing them is to alert you to their existence (see Taylor & Russell, 1939) and to put their importance in the context of evaluating the meaning of a validity coefficient.

Utility analysis is a third method of interpreting the meaning of a validity coefficient, and it can be seen as expanding on the logic of the BESD and the Taylor-Russell tables. Utility analysis frames validity in terms of a cost versus benefit analysis of test use. That is, "Is a test worth using, do the gains from using it outweigh the costs?" (Vance & Colella, 1990, p. 124). Although a full discussion of utility analysis is beyond the scope of this section, we will provide a brief overview.

For a utility analysis, researchers assign monetary values to various aspects of the testing and decision-making process. First, they must estimate the monetary benefit of using the test to make decisions, as opposed to alternative decision-making tools. For example, they might gauge the monetary benefit of hiring employees based partially on test scores as opposed to hiring employees without the test scores. Note that the logic of the Taylor-Russell tables provides some insight into this issue—for example, showing the proportion of applicants selected on the basis of the test who would show successful job performance. Researchers might then estimate the monetary impact of hiring a specific number of people who show successful job performance. Second, researchers must estimate the monetary cost of implementing the testing procedure as part of the decision-making process. The testing procedure might incur costs such as purchasing and scoring the test(s), training decision makers in the interpretation and use of test scores, and time spent by test takers and decision makers in using the test(s). As an outcome of a utility analysis, researchers can evaluate whether the monetary benefits of test use (which, again, are affected by the ability of the test to predict important outcomes) outweigh the potential costs associated with test use.

An analysis of test sensitivity and test specificity is a fourth approach to evaluating the practical effects of using a specific test. Particularly useful for tests that are designed to detect a categorical difference, a test can be evaluated in terms of its ability to produce correct identifications of the categorical difference. For example, a test might be intended to help diagnose the presence versus absence of a specific psychological disorder. In such a case, there are four possible outcomes of the diagnosis, as shown in Table 9.7:

1. True positive—The test leads test users to a correct identification of a test taker who truly has the disorder.

2. True negative—The test leads test users to a correct identification of a test taker who truly does not have the disorder.

3. False positive—The test leads test users to mistakenly identify an individual as having the disorder (when the individual truly does not have the disorder).

4. False negative—The test leads test users to mistakenly identify an individual as not having the disorder (when the individual truly does have the disorder).

Obviously, test users would like a test to produce many correct identifications and very few incorrect identifications.

Sensitivity and specificity are values that summarize the proportion of identifications that are correct. As shown in Table 9.7, sensitivity reflects the ability of a test to identify individuals who have the disorder, and sensitivity reflects the ability of a test to identify individuals who do not have the disorder. More technically, sensitivity reflects the probability that someone who has the disorder will be identified correctly by the test, and specificity reflects the probability that someone who does not have the disorder will be identified correctly by the test. In practice, researchers and test users could never truly know who has the disorder, but sensitivity and specificity are estimated through research that uses a highly trusted standard for gauging whether an individual has the disorder.

Table 9.7 Example of Sensitivity and Specificity

		In Reality, Disorder Is			
		Present	*Absent*		
Test Results Indicate That Disorder Is	*Present*	80 True positive	120 False positive	All with positive test 200	Positive predictive value 80/200 = .40
	Absent	20 False negative	780 True negative	All with negative test 800	Negative predictive value 780/800 = .975
		All with disorder 100	All without disorder 900	Everyone = 1,000	
		Sensitivity 80/100 = .80	Specificity 780/900 = .87	Base rate (prevalence, pretest probability) = 100/1,000 = .10	

In sum, tools such as the BESD, the Taylor-Russell tables, utility analysis, and sensitivity/specificity allow test users and test evaluators to more concretely illustrate the implications of a particular validity coefficient and the use of a given test. Such procedures are clearly important and useful when a test is tied closely to a specific outcome, characteristics, or decision.

Guidelines or Norms for a Field

Yet another way in which validity correlations should be evaluated is in the context of a particular area of research or application. Different areas of science might have different norms for the size of the associations that are typically found. Some areas have greater experimental control over their variables than other areas. Some areas have more precise measurement techniques than others. Some areas may have more complex phenomena, in terms of multidetermination, than others. Such differences affect the magnitude of results obtained in research.

Researchers in the physical sciences might commonly discover associations that most psychologists and other behavioral scientists would consider incredibly strong. For example, a 2000 study examined the association between the mass of black holes at the center of a galaxy and the average velocity of stars at the edge of the galaxies (Gebhardt et al., 2000). This study included approximately 26 galaxies (the "subjects" in this study), and two variables were measured for each galaxy. One variable was the size of the black hole at the center of the galaxy, and the other was the velocity of the stars that orbit on the edge of the galaxy. Analyses revealed a correlation of .93 between the two variables. Such a high correlation is rarely, if ever, found with real data in psychology. Similarly, Cohen (1988) notes that researchers in the field of classical mechanics often account for 99% of the variance in their dependent variables.

In psychology, Jacob Cohen is often cited as providing rough guidelines for interpreting correlations as small, medium, or large associations. According to Cohen's (1988) guidelines for the interpretation of correlations, correlations of .10 are considered small, correlations of .30 are considered medium, and correlations of .50 are considered large (note that Cohen provides different guidelines for interpreting other effect sizes, such as *d*). More recently, Hemphill (2003) conducted a review of several large studies and suggests that a more appropriate set of guidelines would cite correlations below .20 as small, correlations between .20 and .30 as medium, and correlations greater than .30 as large.

Even within the field of psychology, different areas of research are likely to have different expectations for their effect sizes. For example, Hemphill's (2003) guidelines derive from studies in psychological assessment and treatment. The degree to which his guidelines are appropriate for other areas of psychology or the behavioral sciences in general is unknown. Similarly, Cohen (1988) acknowledges that his guidelines "may be biased in a 'soft' direction—i.e., towards personality-social psychology, sociology, and cultural anthropology and away from experimental and physiological psychology" (p. 79). In sum, the interpretation of validity coefficients, as with any measure of association, needs to be done with regard to the particular field.

Statistical Significance

If you read a study that revealed a predictive validity coefficient of .55 for the SAT, would you interpret the result as providing evidence of convergent validity? Using the BESD procedure, a correlation of this size would produce a success rate of nearly 80%, in terms of admitting students with high SAT scores into college. However, what if you found out that the study included only 20 participants? Would this change your opinion of the study? If so, how? What if you found out that the study included 200 participants? Would this improve your opinion of the study? In what way is this a better study?

Earlier in this chapter, we mentioned a real study of the predictive validity of the SAT. This was a large study, including more than 100,000 students from 25 colleges. What is the benefit of such a large study? Is it necessary to have such a large study? As you might know, most studies in psychology, including most validation studies, include much smaller samples—typically a few hundred participants at the most. What, if anything, is lost by having samples of this size?

Statistical significance is the final consideration we will discuss in evaluating evidence of convergent and discriminant quality. Statistical significance is an important part of what is called *inferential statistics,* which are procedures designed to help us make inferences about populations. We will take a moment to explain a few basic issues in inferential statistics, and then we will consider their role in interpreting validity evidence.

Most studies include a relatively small sample of participants. These participants provide the data that are analyzed and serve as the basis for interpretations and conclusions. But researchers usually want to make conclusions about people beyond the few in their particular study. Indeed, researchers usually assume that the participants in their studies represent a random sample from a larger population of people. The 20, 200, or 100,000 people who happen to be included in a SAT study are assumed to represent all students who might take the SAT and attend college.

Because the sample of participants in a study is assumed to represent a larger population, researchers further assume that the participants' data represent (more or less) data that would be collected from the entire population. Thus, they use the data from the sample to make inferences about the population that the sample represents. For example, researchers who find a predictive validity coefficient of .55 for the SAT would like to believe that their results apply to more than the 20, 200, or 100,000 people who happened to be included in their study.

However, researchers are aware that making inferences from a relatively small sample to a larger population is an uncertain exercise. For example, just because data from 20 participants reveal a predictive validity correlation of .55 for the SAT, should we have great confidence that the SAT has predictive validity in the entire population of participants who might take the SAT? It is possible that the sample of 20 people does not represent the entire population of students who might take the SAT. Therefore, it is possible that the predictive validity results found in the sample do not represent the actual predictive validity in the entire population.

Researchers use inferential statistics to help gauge the confidence that they should have when making inferences from a sample to a population. Researchers compute inferential statistics alongside statistics such as correlations to help them gauge the representativeness of the correlation found in the sample's data. Roughly stated, if a result is deemed "statistically significant," then researchers are fairly confident that the sample's result is representative of the population. For example, if a study reports a statistically significant positive predictive validity correlation for the SAT, then researchers feel confident in concluding that SAT scores are in fact positively associated with college GPAs in the population from which the study's sample was drawn. On the other hand, if a result is deemed to be statistically nonsignificant, then researchers are not confident that the sample's result represents the population. For example, if a study reports a statistically nonsignificant positive predictive validity correlation for the SAT, then researchers conclude that the positive correlation in the sample might have been a fluke finding caused purely by chance. That is, they are not willing to conclude that SAT scores are positively associated with college GPAs in the population from which the study's sample was drawn.

With this background in mind, you are probably not surprised to learn that many researchers place great emphasis on statistical significance. Many researchers tend to view statistically significant results as "real" and worth paying attention to, and they view nonsignificant results either as meaningless or as indicating a lack of association in the population. Although these views are not entirely accurate, they seem to be common.

Thus, the size of a validity coefficient is only part of the picture in evaluating the evidence for or against construct validity. In addition to knowing and interpreting the validity coefficient itself (e.g., is it small, medium, or large?), test developers, test users, and test evaluators usually want to know whether the validity coefficient is statistically significant. When evaluating convergent validity evidence, researchers expect to find validity coefficients that are statistically significant. When evaluating discriminant validity evidence, researchers expect to find validity coefficients that are nonsignificant (i.e., indicating that the test might not be correlated with the criterion in the population).

Because statistical significance is often such an important part of the interpretative process, we believe that you should have a basic understanding of the issue being addressed and the factors affecting statistical significance. As applied to the typical case of a validity coefficient, statistical significance addresses a single question—do we believe that there is a nonzero validity correlation in the population from which the sample was drawn?

Note that this is a "yes or no" question. The statistical significance process leads to a dichotomous decision—researchers either conclude that there probably is an association between a test and criterion in the population, or they conclude that there might not be an association between the test and criterion in the population. Again, when evaluating convergent validity, researchers would like to conclude that there is an association between test and criterion in the population, so they hope to find results that are statistically significant. When evaluating discriminant validity, researchers would like to conclude that there is no (or a small) association between test and criterion, so they hope to find results that are nonsignificant. In fact, Campbell and Fiske (1959) included statistical significance as a key part of interpreting the results of an MTMMM analysis.

A more sophisticated version of the basic question is, Are the results in the sample compelling enough to make us confident that there is a nonzero correlation in the population from which the sample was drawn? This highlights the notion of confidence, and it generates two subquestions outlining the factors affecting statistical significance. One question is, *How confident are we* that there is a nonzero validity correlation in the population from which the sample was drawn? The second question is, *Are we confident enough* to actually conclude that there is a nonzero correlation in the population from which the sample was drawn?

There are two factors affecting the amount of confidence that there is a nonzero correlation in the population—the size of the correlation in the sample's data and the size of the sample. First consider the fact that larger sample correlations increase the confidence that the population correlation is not zero. If the correlation between SAT scores and GPA is literally zero in a population, then what correlation would we be likely to find in a sample of people drawn from that population? Even if the correlation in the population is exactly .00, we might not be very surprised to find a small correlation of .07 in a sample. Such a small correlation is only slightly different from the population correlation. We might not even be too surprised to find a correlation of .15 in a sample. Going farther, we might not be shocked to find a somewhat larger correlation of, say, .30 in a sample, even if the sample comes from a population in which the correlation is zero. Such a result (a correlation of .30) is not likely, but it is not impossible. In fact, it is possible that a very strong correlation (e.g., a correlation of .89) could occur in a sample, even if the sample comes from a population in which the correlation is actually zero. In short, relatively large correlations are unlikely to occur (though not impossible) in a sample's data if the sample is drawn from a population in which the correlation is zero. Therefore, larger correlations in the sample's data increase our confidence that the population correlation is not zero. Consequently, larger correlations in the sample data increase the likelihood that the correlation will be considered statistically significant.

Sample size is the second factor affecting the amount of confidence that there is a nonzero correlation in the population. All else being equal, larger samples increase confidence when making inferences about the population. Imagine that you hear about a study reporting a correlation of .30 between SAT scores and college GPA. If you knew that this study included only 20 participants, then how confident would you be in concluding that there is a positive correlation between SAT scores and college GPA among *all* students who could take the SAT? What if you knew that this study included 200 participants or 100,000 participants? Obviously, larger sample sizes should make us more confident when making conclusions about a population.

In sum, the size of the correlation and the size of the sample affect our confidence in concluding that there is a nonzero correlation in the population. The precise statistical equations are beyond the scope of this discussion, but in general, larger correlations and larger samples increase our confidence that the population in the correlation is not zero. Thus, larger correlations and larger samples increase the likelihood that the results of the validity study will be statistically significant. An equation (based on Rosenthal, Rosnow, & Rubin, 2000) summarizes the issue:

However, for results to be deemed statistically significant, we must have a specific level of confidence that the population correlation is not zero.

Confidence that test is correlated with criterion *in the population*	=	Size of validity coefficient in the sample	×	Size of sample

Thus, the second question regarding statistical significance is, *Are we confident enough* to actually conclude that there is a nonzero correlation in the population from which the sample was drawn? Large correlations and large sample sizes increase our confidence, but we must ask if the results of a particular study make us confident *enough* to deem results statistically significant. We must set a specific level of confidence as a cutoff point that must be met before we conclude that the population correlation is not zero. By tradition, most behavioral researchers use a 95% confidence level as the cutoff point for declaring results to be statistically significant. Put another way, most behavioral researchers are willing to declare results statistically significant if they find that there is only a 5% chance of being wrong (i.e., a probability of .05). This cutoff is the "alpha level" of a study. If our inferential statistics surpass the alpha level, then we are confident enough to conclude that there is a nonzero validity correlation in the population from which the sample was drawn.

As mentioned earlier, statistical significance is an important issue in interpreting evidence for convergent and discriminant validity. The fact that statistical significance is affected by sample size, effect size (i.e., the size of the validity coefficient in the sample), and the alpha level is an extremely important point. These issues should be considered when interpreting inferential statistics. The results of a validity study can be statistically significant even if the validity correlation is quite small. This could occur if the size of the sample in the validity study was sufficiently large. Similarly, the results of a validity study can be nonsignificant, even if the validity correlation is quite large. This could occur if the size of the sample in the validity study was quite small.

We mentioned earlier that most researchers would hope to find convergent correlations that are statistically significant, and they would hope to find discriminant correlations that are nonsignificant. But what are the implications of finding a convergent validity correlation that is nonsignificant? The typical interpretation would be that the test in question has weak convergent validity (i.e., that the convergent correlation might well be zero in the population). However, such a result should be interpreted with regard to the size of the correlation and the size of the sample. A nonsignificant convergent validity correlation could occur because the correlation is small or because the sample is small. If the correlation is small, then this is certainly evidence against the convergent validity of the test. However, if the correlation is moderate to large in size but the sample is small, then the results might not indicate poor convergent validity. Instead, the results could indicate a poorly conceived study. If a study included a sample that was too small, then perhaps a larger study should be conducted before making any conclusions about construct validity.

Similarly, what are the implications of finding a discriminant validity correlation that is statistically significant? The typical interpretation would be that the test in question has weak discriminant validity (i.e., that the discriminant correlation is probably not zero in the population). Again, such a result should be interpreted with regard to the size of the correlation and the size of the sample. A significant discriminant validity correlation could occur because the correlation is large or because the sample is large. If the correlation is large, then this is certainly evidence against the discriminant validity of the test. However, if the correlation is small but the sample is quite large, then the results might not indicate poor discriminant validity. It is possible that a correlation of only .10, .06, or even smaller could be statistically significant if the sample is large enough (say in the thousands of participants). In such cases, the statistical significance is almost meaningless and should probably be ignored.

In sum, statistical significance is an important but tricky concept as it is applied to validity evidence. Although it plays a legitimate role in the interpretation of convergent and discriminant validity coefficients, it should be treated with some caution. As a rule, convergent correlations should be statistically significant, and discriminant validity correlation should be nonsignificant. However, this general rule should be applied with an awareness of other factors. A sophisticated understanding of statistical significance reveals that the size of the sample and the size of the convergent and discriminant validity correlations both determine significance. Thus, a nonsignificant convergent correlation could reflect the fact that the study had an inadequate sample size, and a significant discriminant correlation could reflect the fact that the study had an extremely large sample size.

Summary

Convergent and discriminant evidence is key to the empirical evaluation of test validity, and this chapter presents issues related to the estimation and evaluation of these important forms of validity evidence. We began by describing four methods that have been used to estimate and gauge the degree of convergence and discrimination among tests (e.g., multitrait-multimethod matrices). We then discussed seven factors that can affect the size of validity coefficients (e.g., measurement error, relative proportions, method variance). Finally, we presented four important issues that should be considered when judging the meaning and implications of validity coefficients (e.g., variance explained, statistical significance, practical importance). Awareness of the issues described in this chapter can provide a more sophisticated and informed perspective on the meaning and evaluation of test validity.

Suggested Readings

This is a discussion of the interpretation of effect sizes:

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97,* 129–133.

This is a classic article, in which the multitrait-multimethod matrix is presented for the first time:

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait multimethod matrix. *Psychological Bulletin, 56,* 81–104.

This presents the Taylor-Russell tables:

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565–578.

This article presents the logic and computation details of the quantifying construct validity procedure:

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

This presents an overview of the concept of statistical power, which is an important issue in evaluating the statistical significance of validity coefficients:

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

This paper presents a factor-analytic approach to the examination of MTMMM data:

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait multimethod data. *Applied Psychological Measurement, 9,* 1–26.