Statistical and Clinical Interpretation Guidelines for School Neuropsychological Assessment

10

W. Joel Schneider

Policy, Organization, and Leadership Studies, College of Education and Human Development, Temple University, Philadelphia, PA, USA

Mastery of psychometrics—the science of measurement—gives assessment professionals a powerful and flexible toolkit for answering important questions about individuals. Without a strong knowledge of psychometrics, one can do excellent work, but in limited domains. An understanding of these principles can mark the difference between a competent professional and a true expert. This chapter is aimed at practitioners who want to expand their professional toolkit for helping the people they assess.

This chapter distinguishes between the kind of psychometric expertise that is needed to become a test developer and the kind of psychometric savvy one needs to extract useful information from test scores. It is possible that a test developer may have relied on the most sophisticated psychometric procedures available, yet the end user need not have a complete understanding of the subtle complexities of test design. However, one does need to understand the scores being used and how to correctly interpret them psychometrically. In many domains (e.g., learning disorder diagnosis, inference of memory processing dysfunction, estimation of the extent of brain injury), the evaluator needs to think globally, flexibly, and interactively about test scores to arrive at valid conclusions about underlying psychological processes.

What follows is not a primer on "test developer psychometrics," but a brief overview of basic psychometric principles followed by a set of procedures that may be useful in answering practical questions in individual assessment cases. To avoid repeated citations, it should be noted that this chapter draws heavily on many authoritative sources (e.g., Cohen et al., 2003; Crocker & Algina, 2006; Furr, 2017; McDonald, 1999; Nunnally, 1967; Raykov & Marcoulides, 2011).

Basic Statistics to Describe Unique Characteristics of Individuals

Typical readers of this chapter have taken statistics courses in which they learned how to describe and analyze data sets collected by researchers. In such courses, we have data from individuals, and we hope to learn something about populations and the theories that explain their behavior. Here, our task is the reverse. We have data-based theories about populations, and we hope to learn something about an individual we have assessed. Instead of using sample statistics like means, standard deviations, and correlations to describe what is typical in populations, we use population means, standard deviations, and correlations as benchmarks to describe the unique characteristics of individuals. Doing so requires a different mindset about statistics and learning some additional conceptual tools.

Best Practices in School Neuropsychology: Guidelines for Effective Practice, Assessment, and Evidence-Based Intervention, Second Edition. Edited by Daniel C. Miller, Denise E. Maricle, Christopher L. Bedford, and Julie A. Gettman.

Reliable and Valid Measurement

If scores from a test differ dramatically each time the test is given to a person, no single score from that test can be trusted to be accurate. By contrast, if evidence and experience tell us that a test yields roughly the same score each time the same person takes it, then we are justified in giving the test just once. *Test reliability* refers to consistency of measurement. For trait-like characteristics such as reading ability and math skill, we expect measurements of the same person to be stable from one measurement to the next, at least over the short term. Unfortunately, some variables in psychology are not expected to be stable (e.g., mood), yet we still would like to know if our measurements of such variables are reliable. This difficult problem can be solved by invoking an ingenious fiction: the distinction between *construct scores* (Borsboom & Mellenbergh, 2002; Lord et al., 1968) and *true scores* (Spearman, 1904).

Construct Scores and Test Validity

A *construct* is a theoretical entity that explains a wide array of behaviors. For example, if a person has persistently sad mood, relentless negative thoughts about the self, hypersomnia, low energy, and reduced appetite, we would infer that the person has high levels of the theoretical construct we call *depression*. Every test we give is intended to measure one or more constructs, including constructs related to ability, personality, psychopathology, interests, attitudes, moods, preferences, beliefs, and opinions. A test score helps us know how high or low a person is on the theoretical construct's number line of possible scores.

If we had a perfect measurement procedure that measured the construct perfectly and without fail, we would know each person's *construct score* (see Figure 10.1). From these perfect measurements, we could plot the population distribution and evaluate each person's relative standing on the construct. With perfect measurement, the only reason observed scores would differ from one person to the next is that each person has a different construct score.

Unfortunately, perfect measures do not exist. Each measurement is influenced directly by its intended construct and also by any number of irrelevant influences. We would like to be able to quantify how much an observed score reflects the intended construct and how much it reflects other influences. *Valid variance* refers to variability in test scores due to variability in the construct scores (Borsboom et al., 2004). A good test score consists mostly of valid variance. In the observed score in Figure 10.1, the score's validity coefficient is 0.75, meaning that 75% of the observed score's variability consists of valid variance. The remaining 25% of the variability in the score is due to influences other than the construct and is thus not valid.

Reliability and Measurement Error

All valid variance is reliable, but not all reliable variance is valid. *Reliable variance* refers to variability in test scores due to stable influences. In a good test, most reliable variance is valid variance. Unfortunately, test scores always have influences—however small—that are stable but unrelated to the construct we are trying to measure. For example, if a math test requires reading skills in English to understand the questions, the math test scores will reflect not only differences in math skill but also differences in reading ability in English. Thus, the test will tend to underestimate the math skills of students with lower reading skills or students with relative unfamiliarity with English.

The term *unsystematic measurement error* refers to short-lived, hard-to-anticipate influences on test scores that differ from person to person and moment to moment, such as fluctuations in motivation, attention, and energy. It also includes test score influences that are unlikely to be repeated, like examiner error and disruptive noises in the testing environment.

As you may have anticipated from the term *unsystematic*, there is a second kind of measurement error. *Systematic measurement errors* refer to stable, construct-irrelevant influences on test scores that are likely to occur each time one completes the measure. These include imperfections in the test design (e.g., confusing instructions) and response biases in the examinees (e.g., younger children often give extreme responses on questionnaires).

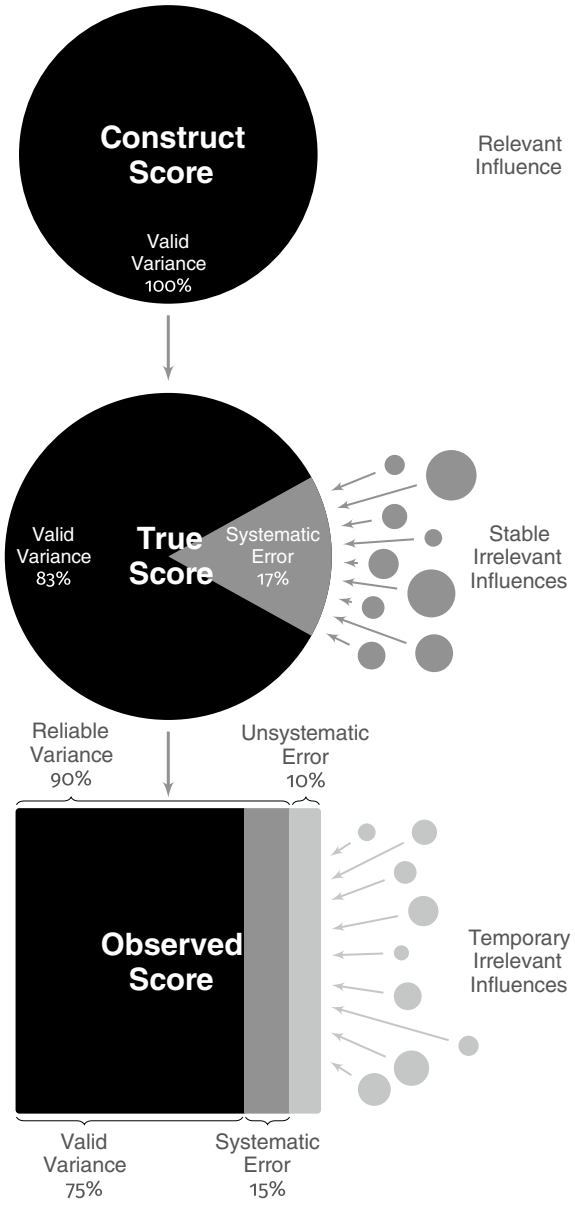

FIGURE 10.1 Construct scores, true scores, and measurement error.

Most of the time when scholars refer to *measurement error*, they mean unsystematic measurement error. For the remainder of this chapter, *error* or *measurement error* refers to unsystematic measurement error, not systematic error.

True Scores Are Not Construct Scores

The *true score* is the score that we would obtain on a test if there were no (unsystematic) measurement error. That is, every observed score X is the sum of its true score T and its unsystematic measurement error E:

$$X = T + E$$

Despite its name, the true score does not necessarily refer to the truth (i.e., the construct score). True scores are the sum of all reliable influences, including the construct scores and systematic measurement error. Some true scores consist of more systematic measurement error than valid variance. Thus, it is possible to estimate a person's true score on a test, whether the test is valid or not. Good tests, mediocre tests, and even ridiculous, ill-conceived, and thoroughly awful tests have true scores. That is, true scores are consistent but not necessarily valid.

Carryover Effects

Imagine that we could give a test to a child again and again without earlier administrations of the test influencing subsequent administrations. *Carryover effects* occur when previous testing influences subsequent testing. Carryover effects include *practice effects*, such as when a child learns how to solve a problem more quickly during testing and therefore performs better the next time the test is given. Carryover effects can also negatively influence performance, such as when a child becomes bored with a test and no longer tries hard on subsequent administrations. To eliminate carryover effects entirely, we would have to imagine that we could rewind time repeatedly such that the person had no memory of being tested before. Of course, the idea that we could rewind time is preposterous. Yet, some preposterous ideas are rather useful. It allows us to think about reliability without worrying about THE TRUTH, which, as far as we mortals are concerned, can only be approximated.

The average of a person's potential score on a particular test is the *true score* for that test. For example, in Figure 10.2, the population mean is 100, and the standard deviation is 15. The gray points in the figure represent all the possible scores a particular child could obtain in various settings and situations. These potential scores have a mean of 85, which is the child's true score for this test. The standard deviation of these scores is the standard error of measurement (SEM) ($\sigma_e = 5$). It is unlikely that the child will score lower than 70 or greater than 100 on this test. Most likely, the child will score near 85 ± 5.

The term *true score* was coined before the distinction between systematic and unsystematic measurement error was clearly understood (Borsboom & Mellenbergh, 2002). If we could rewind time and rename the true score, we might call it something less likely to be misunderstood (e.g., *personal long-term average*). Unfortunately, we have been calling it the *true score* for over a century, and it is probably too late for change.

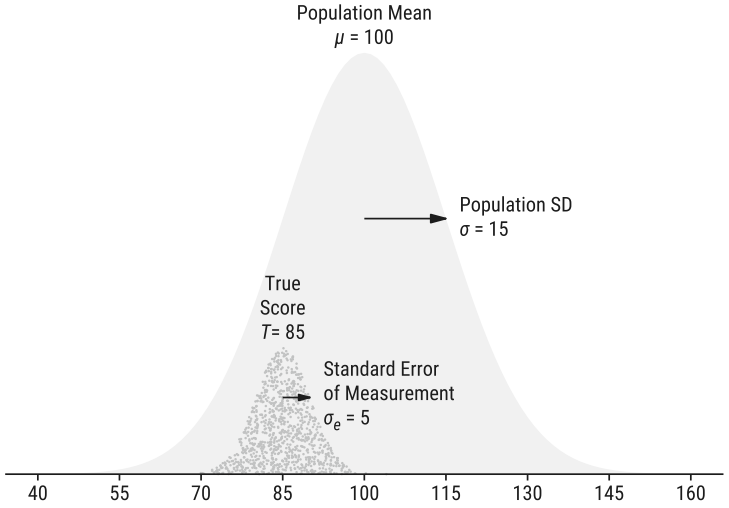

FIGURE 10.2 A true score is the average of repeated measurements from the same person.

The Standard Error of Measurement

In most classical test theory models, it is assumed that unsystematic measurement errors have the same distribution for all members of a particular population. The standard deviation of all the measurement errors is called the *standard error of measurement*. However, we know that some people produce more consistent scores than others. For example, very young children are famously fickle when it comes to giving their best performance. For this reason (and others), reliability is estimated separately for different age groups. However, even after accounting for age, sex, race, and other demographic distinctions, some children are more sensitive than others to situational factors and their own emotional fluctuations. For example, children with attention-deficit/hyperactivity disorder tend to have a wide variation in their performance on attention-demanding tests, with scores ranging from average or better to severely impaired (Castellanos et al., 2005; Johnson et al., 2007; Klein et al., 2006). It is for this reason that continuous performance tests measure not only a child's level of performance but also their variability of performance.

The size of the typical measurement error differs from person to person, often for reasons that have little to do with the person and everything to do with the test. For example, a screening test for identifying students with academic deficits needs to have only a few easy items. Such a test can reliably distinguish between students with and without academic skill deficits, but it cannot reliably distinguish between students with average and high skills. Thus, low scores on the test have a smaller measurement error than high scores on the test have.

Reliability Coefficients

Technically, the question "How reliable is the test?" is ill posed because reliability is not guaranteed to be the same for all people taking a particular test. Reliability is a joint property of a particular test given to a particular person in a particular situation (Thompson & Vacha-Haase, 2000). It is therefore more appropriate to talk about the reliability of specific test scores than to talk about the reliability of whole tests.

Nevertheless, it is useful to know how reliable test scores are on average in a particular population. A reliability coefficient is a summary statistic. It does not tell us the reliability of a particular score but does give us some idea as to what level of reliability is typical for scores in a particular population.

The classical test theory definition of the reliability coefficient (ρ_{XX}) is the proportion of variance in a score that is due to true score variance:

$$\rho_{XX} = \frac{\sigma_T^2}{\sigma_X^2}$$

The variance of the observed score in the denominator of this equation is easily estimated, but true score variance in the numerator must be estimated indirectly. There are several indirect methods of estimating reliability, and they do not always give the same result, which means that we have to pay careful attention to the assumptions that underlie each method.

Retest Reliability

The *retest reliability coefficient* measures the typical stability of scores for a particular test over a specified time interval. The retest reliability coefficient is the correlation between scores on a test that have been given to the same sample twice.

A test does not have just one retest reliability coefficient, because the retest reliability can change depending on the interval between the two tests. In general, the longer the interval, the smaller the retest reliability coefficient. Why? Very few things about human beings stay exactly the same over time. Some individual difference constructs are fairly stable (e.g., intelligence, personality), and we

can call them *traits*. Some constructs are expected to change often (e.g., mood), and we can call them *states*. However, it is well known that even fairly stable constructs fluctuate a little. To the degree that a construct is unstable, the retest reliability coefficient will underestimate the average reliability of the scores.

A second reason that retest reliability coefficients are sometimes inaccurate is that the act of measurement can alter subsequent measurements. Large carryover effects can increase or decrease the retest reliability coefficient.

Alternate-Form Reliability

To reduce carryover effects, instead of administering the same test twice, an alternate form of a test can be administered on the second administration. It is common for academic achievement tests to have alternate forms (e.g., *Woodcock-Johnson IV* [WJ IV] *Tests of Academic Achievement, Forms A, B, and C*; Schrank et al., 2014). The *alternate-form reliability coefficient* is the correlation between the two tests. The benefit of reducing carryover effects is dearly bought: it requires the construction and norming of a completely new test.

Split-Half Reliability

An alternate-form reliability coefficient can be had "on the cheap" if one is willing to cut one's test in half. If the items of a test are divided in a sensible manner, the correlation between the totals of the two halves is an estimate of the reliability of each half (e.g., two-part speeded tests like the *Comprehensive Test of Phonological Processing, Second Edition* [CTOPP-2; Wagner et al., 2013] rapid naming tests). Unfortunately, we are uninterested in the reliability coefficient of half a test—we want to know the reliability of the whole test. Fortunately, there is a statistical correction that can be applied such that the reliability of the halves can estimate the reliability of the whole.

Spearman-Brown Prophecy Formula

Discovered independently by both Spearman (1910) and Brown (1910) at about the same time, this formula estimates how reliable a test becomes when we increase its length by adding parallel items (i.e., items similar to the existing items):

$$\rho_{CC} = \frac{k\rho_{XX}}{1+(k-1)\rho_{XX}}$$

where

ρ_{CC} = Reliability of the extended test

ρ_{XX} = Reliability of current test

k = Ratio of the number of extended test items to the number of current test items

Imagine that the split-half reliability coefficient of a test is 0.80. We would like to know the reliability of the whole test, which has twice as many items. Applying the Spearman-Brown Prophecy Formula, where $k = 2$ (because we are doubling the number of test items), we see that:

$$\rho_{CC} = \frac{2\times 0.80}{1+(2-1)\times 0.80} \approx 0.89$$

If a reliability coefficient of 0.89 is not high enough, the Spearman-Brown prophecy formula can be rearranged to tell how many more parallel test items are needed to achieve a reliability coefficient that is sufficiently high. For example, suppose that a test needs to have a reliability coefficient of 0.95. The Spearman-Brown Prophecy formula rearranged is:

$$k = \frac{\rho_{CC}(1-\rho_{XX})}{\rho_{XX}(1-\rho_{CC})}$$

Applying the rearranged formula, we see that:

$$k = \frac{0.95\left(1-0.89\right)}{0.89\left(1-0.95\right)} \approx 2.35$$

If our original test has 10 items, we will need about 24 items total ($10 \times 2.35 = 23.5$) to increase the reliability coefficient to 0.95.

Coefficients Alpha and Omega as Measures of Internal Consistency

The split-half coefficient is almost never used in practice because its value changes depending on how the test is split. To address this problem, alternate measures of internal consistency were developed that give consistent answers. The most popular measure of internal consistency is coefficient α, which is the average split-half coefficient after the Spearman-Brown correction has been applied (Cronbach, 1951). That is, if we split a test in half every way possible and calculate the split-half reliability coefficient, applying the Spearman-Brown correction each time, the average of all our calculations will be coefficient α, often referred to as *Cronbach's alpha*.

Coefficient α is a good estimate of reliability but slightly underestimates reliability when test items are unequally related to the construct they are intended to measure. Increasingly, scholars use coefficient omega-total (McDonald, 1999) as a (slightly) more accurate replacement for coefficient α (Dunn et al., 2014).

How High Should a Reliability Coefficient Be?

How high does a test's reliability coefficient need to be before we can recommend its use? It would be nice to give a good and simple answer to this question, but no simple answer is very good. Reliability coefficients have a possible range from 0 to 1, and there are no hard boundaries between coefficients that are "good enough" and coefficients that are "too low." You sometimes hear about convenient threshold values like "A reliability coefficient of 0.80 is high enough for a test used as a screener, 0.90 is high enough for use in low-stakes decisions, and 0.95 is high enough for use in high-stakes decisions." Such recommendations are well-intentioned but ought not to be taken too seriously. The reliability of a score and the reliability of a decision based on the score are not quite the same thing. Under some conditions, highly reliable decisions can be based on scores with rather modest reliability coefficients, and unreliable decisions can be made with scores with high reliability. To understand how reliable a decision based on a score is, we can use a statistical concept called a *confidence interval*, which is based on reliability coefficients.

Confidence Intervals

A *confidence interval* is a method for specifying a range in which an unknown quantity is likely to fall. In the assessment of individuals, we often wish to know where a person's true score is likely to be. In this context, a confidence interval can be used to give a sense of the most likely regions the true score falls.

Just as there are many different kinds of reliability coefficients, there are different kinds of confidence intervals (Crawford & Garthwaite, 2002). Two types will be highlighted here. Unfortunately, there is no consensus on what to call them. One is a confidence interval based on the SEM and is centered on the observed score. The other is based on the standard error of the estimate (SEE) and is centered on the estimated true score (Charter & Feldt, 2001).

SEM-Based Confidence Intervals

Although we never truly know a person's true score, we can still estimate how far the observed score is likely to be from the true score. Error scores are the difference between observed scores and true scores:

$$E = X - T$$

The standard deviation of the errors (σ_E) is called the *standard error of measurement* (SEM). It represents a typical distance between the true score and the observed score. Specifically, it is the square root of the average squared distance between the observed scores and the true scores. Using the following equation, you can see that as a reliability coefficient approaches 1, the SEM approaches 0:

$$SEM = \sigma_E = \sigma_X \sqrt{1 - \rho_{XX}}$$

Assuming that the errors are normally distributed, about 68% of observed scores have a true score within the interval of ±1 SEM. If we wish to specify a particular degree of confidence, we multiply a *z*-score by the SEM to create the *margin of error*. The *z*-score associated with the 95% confidence is about 1.96. The upper and lower bounds of the confidence interval are calculated by adding the margin of error to or subtracting it from the observed score:

$$CI_{SEM} = x \pm z\sigma_E$$

Suppose we measure a quantity in many individuals and specify a 95% confidence interval around the score. In about 95% of cases, the confidence interval will contain the individual's true score (See Figure 10.3). This statement is often taken to mean that for every individual, "there is a 95% chance that the true score is contained by the confidence interval." There is some philosophical debate as to whether it is proper to talk about probabilities of any particular confidence interval containing a true score. For purists, the confidence interval either contains the true score or does not. For others, when events have been determined but the outcome is still unknown, it does not seem like too great an error to still talk about the probability that the outcome will be revealed to have turned out one way or another. For a fuller discussion of similar issues, see Crawford et al. (2009).

Although this kind of confidence interval is not particularly hard to calculate, its proper meaning is hard to explain, especially to statistically untrained parents and teachers. It is easy but incorrect to say that "There is a 95% chance that your child's true score is between these two numbers." It is more correct to say that this procedure captures the true score in 95% of children. Thus, the CI_{SEM} is ultimately not really about a particular child's score. Fortunately, there is a different kind of confidence interval that is easier to explain, more relevant to the particular score, and narrower than the CI_{SEM}.

SEE-Based Confidence Intervals

If both observed and true scores were known, we could use regression to predict the true scores from the observed scores and then estimate how far our predictions were from the actual true scores. The SEE is the standard deviation of the prediction errors in a regression analysis.

The confidence interval based on the standard error of the estimate (CI_{SEE}) is narrower than the SEM-based confidence interval (CI_{SEM}) because the SEE is smaller than the SEM:

$$SEE = SEM\sqrt{\rho_{XX}} = \sigma_X \sqrt{\rho_{XX} - \rho_{XX}^2}$$

The CI_{SEE} is calculated like so:

$$CI_{SEE} = \left(X - \mu_X\right)\rho_{XX} + \mu_X \pm z \bullet SEE$$

Suppose that child scores 70 on a test with an index score metric ($\mu_X = 100$, $\sigma_X = 15$). If the reliability coefficient is $\rho_{XX} = 0.80$, the two types of confidence intervals are:

$$\begin{aligned} CI_{SEE} &= \left(X - \mu_X\right)\rho_{XX} + \mu_X \pm z\sigma_X\sqrt{\rho_{XX} - \rho_{XX}^2} \\ &= \left(70 - 100\right)0.8 + 100 \pm 1.96 \cdot 15\sqrt{0.8 - 0.8^2} \\ &= 76 \pm 11.76 \\ &= 64.24 \text{ to } 87.76 \end{aligned}$$

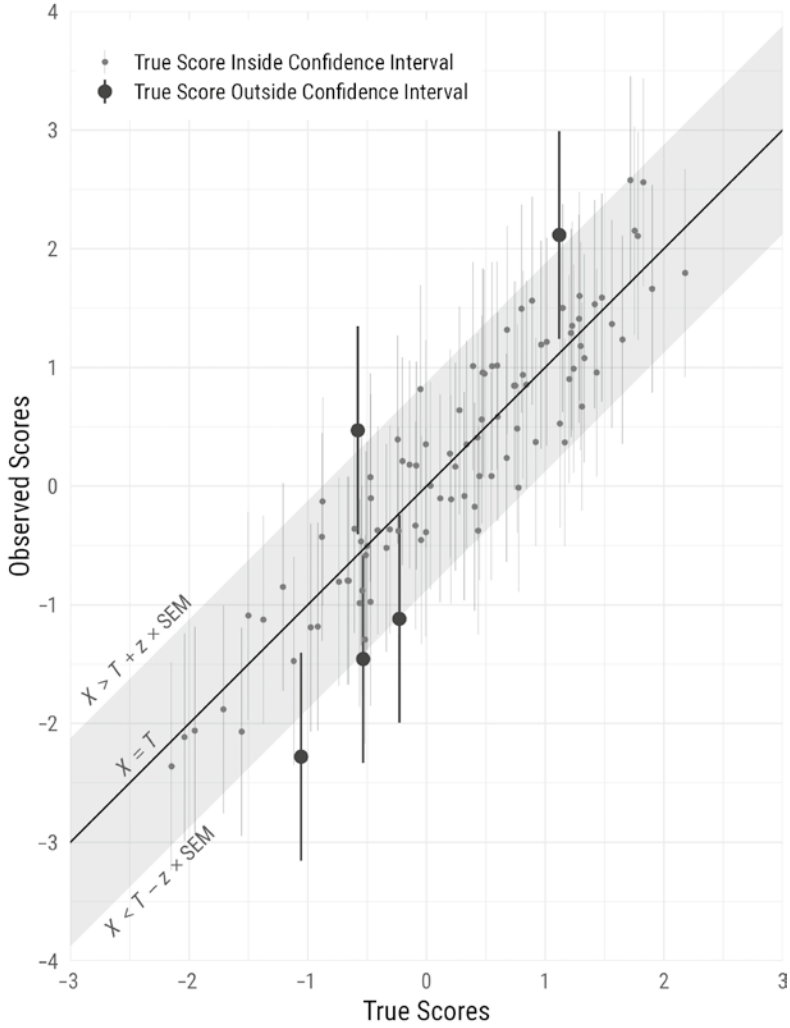

FIGURE 10.3 SEM-based 95% confidence intervals contain the true score 95% of the time.

$$\begin{aligned} CI_{SEM} &= X \pm z\sigma_X\sqrt{1-\rho_{XX}} \\ &= 70 \pm 1.96 \cdot 15\sqrt{1-0.8} \\ &= 70 \pm 13.15 \\ &= 56.85 \text{to} 83.15 \end{aligned}$$

The two procedures give different answers, and yet both are correct 95% of the time. How is it possible that the CI_{SEE} is narrower than the CI_{SEM} yet is equally accurate? To make sense of this seeming impossibility, one must recognize that they are equally accurate answers to different questions. As seen in Figure 10.4, the SEM is the standard deviation of observed scores when the true score is held constant. The SEE is the standard deviation of the true scores when the observed score is held constant. The SEE is narrower than the SEM because true scores have less variability than observed scores. In assessment, we have an observed score in hand and would like to use this information to narrow our search for the true score. For this reason, the CI_{SEE} is the type of confidence interval provided by all major cognitive tests.

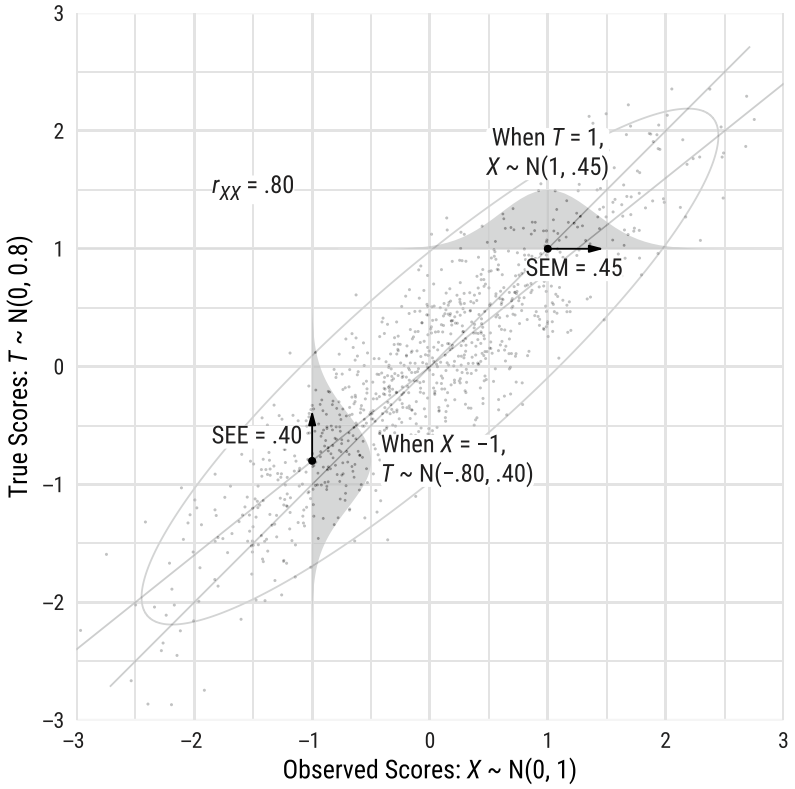

FIGURE 10.4 CI_{SEM} and CI_{SEE} give different results because they answer different questions.

Application of Confidence Intervals in Practice

Confidence intervals around test scores are routinely presented in test protocols and in score reports. If you rely solely on the observed scores, your test interpretations are vulnerable to being overly precise. If the child is retested, the new score is unlikely to be exactly the same. Instead of imagining the scores as fixed values, imagine them dancing around within the range of the confidence interval. If your test interpretation remains roughly the same no matter where the scores move within the confidence intervals, then your interpretation has a far greater probability of being accurate and a far greater probability of being similar to conclusions derived from other evaluations.

Following is an example of how to explain the CI_{SEE} to parents and teachers. It may be helpful to have a normal curve with percentile benchmarks to give the scores context (see Figure 10.5):

> *Some basketball players are better than others at making free throw shots. If we observe a player make 8 out of 10 shots, we can guess that the player's long-term free throw percentage is near 80%. However, this is only a guess, and time will tell how close to reality our guess is. Even if we know a player's true free throw percentage, we cannot know how well the player will perform in any particular game. We only know what is typical for that player.*
>
> *After an evaluation like this, we are in a similar situation. I have a test score that measures your child's ability in Domain X. However, no test is perfect—every score is an estimate, not the precise truth. The same child performs differently on the same test on different days and in different situations. If somehow, we could rewind time over and over and test your child many times, we could average all the scores and get a very accurate estimate. Let's call that accurate number your child's "typical performance."*
>
> *Obviously, we cannot rewind time, and so we may never know your child's true typical performance. However, this test is accurate to a known degree and allows us to make an informed guess*

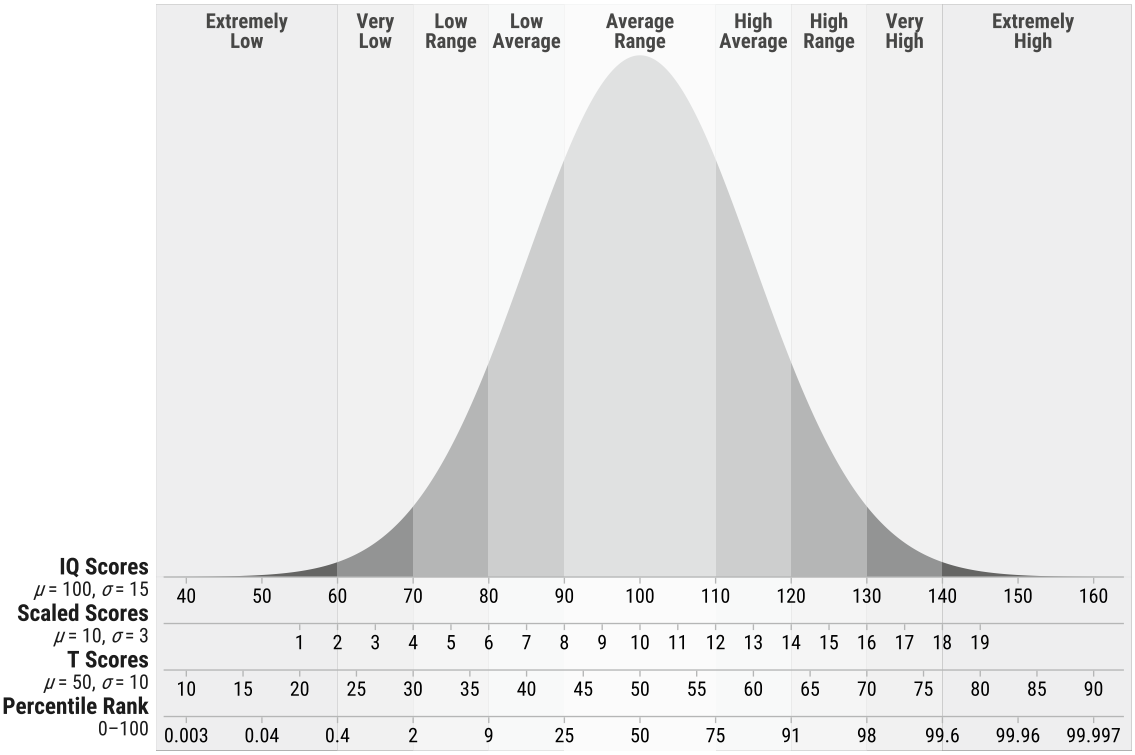

FIGURE 10.5 Standard scores with percentile benchmarks.

as to where the typical performance is likely to be. About 95% of children who score 70 on this test have a typical performance between 64 and 88. These two numbers are the "95% confidence interval." I cannot say for certain where exactly your child would score on average if we could give this test many times without your child getting bored or discouraged; however, somewhere between 64 and 88 is a good guess, most likely in the middle around 76.

Types of Validity

The next section of this chapter will review the various types of validity and how they influence the clinical interpretation of test scores. Figure 10.6 illustrates the relationships among the various types of validity.

Construct Validity

Construct validity refers to the degree to which theory and evidence support the use of a test for a particular purpose (Cronbach & Meehl, 1955; Messick, 1995). Note that this definition implies that validity is neither binary {*valid, not valid*} nor unidimensional; it is a multifaceted/multidimensional phenomenon. There are no thresholds that divide valid tests from tests that are not valid. Instead, scholars consider the totality of evidence that a test measures a particular construct and decide whether the test is appropriate for the purpose they have in mind.

Face Validity

Face validity only barely qualifies as a kind of validity. It refers to a superficial and subjective judgment as to whether a measure appears to measure its intended construct. For instance, we could

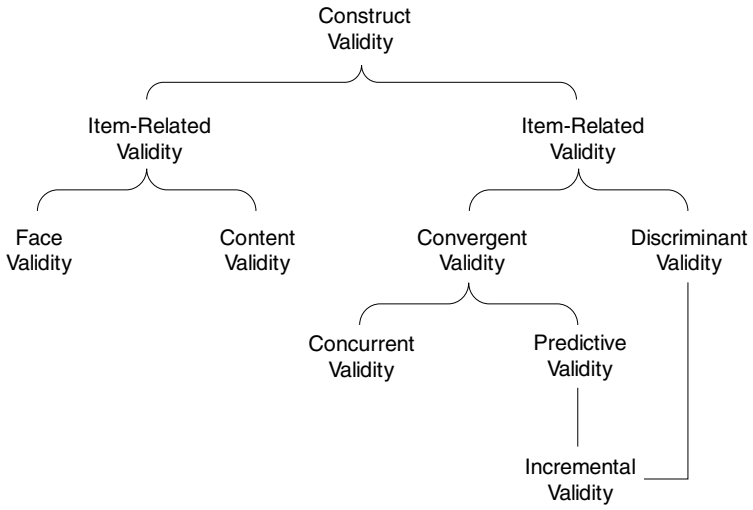

FIGURE 10.6 Different types of validity.

measure sadness in a face-valid way by simply asking children, "Do you feel sad or not sad right now?" A face-valid measure of the ability to say the alphabet is to ask children to recite the alphabet and record how many errors are made. These measures have face validity because they are direct and obvious. They are likely reasonably valid for their intended purposes as well. Note that there is no "face validity coefficient" or measure of face validity. A measure is *face valid* to the degree that people agree it is a straightforward measure of its intended construct.

Some face-valid measures are not likely to be valid at all. For example, asking very young children to estimate how good they are at something is a straightforward method of assessing abilities. However, very young children are unlikely to give accurate self-assessments.

Most tests with proven validity are also face valid. The direct and obvious approach to measurement is often the best method. However, some measures with proven validity are not obviously connected to the construct they are intended to measure. For example, in the Strange Situation, a well-validated method of assessing attachment in infants (Ainsworth et al., 1978), the parent temporarily leaves the child alone in a room with a stranger. One might think that a securely attached infant would trust the parent to return and would not become too alarmed. However, a failure to become visibly alarmed is actually an indication of insecure attachment.

Content Validity

Many, if not most, psychological constructs are too broad to be measured by a single question or test item. *Content validity* refers to how adequately a measure samples the domain of the construct. For example, if we hope to measure reading ability and we only measure decoding, but not fluency or comprehension, we might have adequate coverage of the child's ability to sound out words, but not of reading ability as a whole. Again, there is no such thing as a "content validity coefficient." Experts evaluate a measure and render a judgment as to its content validity.

Criterion-Oriented Validity and the Nomological Network

One of the most important methods of demonstrating test validity is by showing that the measure correlates with other variables according to a pattern predicted by theory. The network of theoretically relevant relationships between the test and other variables is termed the *nomological network* (Cronbach & Meehl, 1955). When first proposed, the evaluation of various relationships in the nomological network was mostly piecemeal and unsystematic. Today, statistical techniques such as

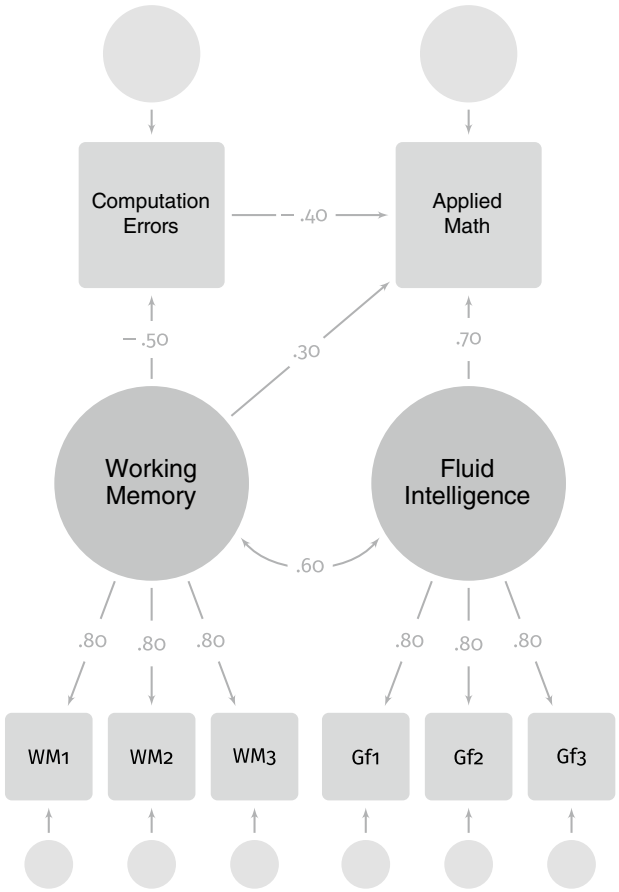

FIGURE 10.7 Construct validity of a battery of working memory and fluid intelligence tests.

structural equation modeling allow for precise evaluations of complex multivariate relationships. For example, in Figure 10.7, many relationships among the constructs are tested simultaneously.

Discriminant Validity

Discriminant validity refers to the idea that tests measuring different constructs should not have higher correlations than predicted by theory. For example, measures of intelligence and conscientiousness are not expected to have more than small correlations. If a new measure of conscientiousness exhibits substantial correlations with intelligence, we would wonder if the measure is contaminated with too much intellectual content.

Convergent Validity

Convergent validity refers to the idea that a test should correlate substantially with other measures of the same construct and also with other theoretically relevant constructs.

If a test correlates highly with other tests measuring the same construct administered on the same occasion, the test has *concurrent validity*. For example, when a brief intelligence test is published, test developers often administer a longer (and well-established) intelligence test to a portion of the standardization sample so that it can be shown that the brief test correlates highly with the more established longer test. In Figure 10.7, three measures of working memory capacity correlate substantially. Likewise, three measures of fluid intelligence show concurrent validity.

Originally, *predictive validity* referred to a test's ability to predict outcomes measured on a subsequent occasion. However, many people use the term *predictive* in a broader sense: Time and sequence are irrelevant, as long as one score is a significant predictor of another in a statistical model. For example, measures of intellectual functioning in adolescents can "predict" aspects of brain functioning measured earlier in childhood (Burgaleta et al., 2014). Obviously, the future cannot cause the past, but it can reveal things about the past that we did not previously know. This broad sense of predictive validity sometimes subsumes concurrent validity, leading to the term *predictive validity* being used synonymously with *convergent validity*.

In Figure 10.7, the fluid intelligence measures have predictive validity in that they predict scores on an applied math test. The working memory measures predict the applied math scores but also "careless" computational errors (i.e., items missed because of lapsed attention rather than a lack of knowledge).

In Figure 10.7, we see that the cognitive tests demonstrate *incremental validity*, which could be termed *discriminant predictive validity*. That is, both sets of cognitive tests provide unique information about the two outcomes. For example, although fluid intelligence is correlated with computation errors, it provides no *incremental* information beyond what we already knew from our working memory tests. In contrast, for applied math performance, the fluid intelligence tests *do* provide information beyond what can be known from the working memory measures. In this case, both fluid intelligence and working memory demonstrate incremental validity in predicting applied math performance. To some degree, they are redundant (i.e., they are correlated), but each provides unique information about applied math performance that the other does not.

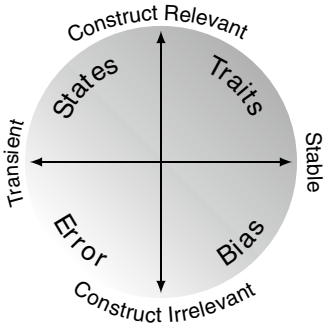

FIGURE 10.8 Different kinds of influences on test scores.

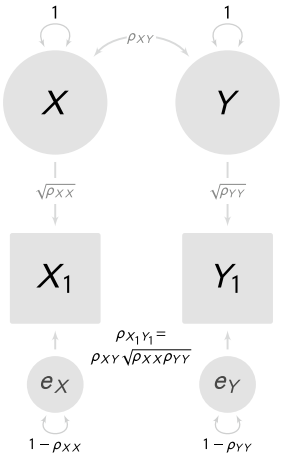

FIGURE 10.9 Attenuation of validity.

Reliability and Validity

Validity and reliability have an interesting relationship. Like reliability, validity is a joint property of persons, situations, and test scores. Unreliable measurement has no hope of being valid. The reverse is not true, however; without validity, reliability is worse than useless—it means that the test gives consistently wrong information. Without validity, a test has no reason to exist.

Figure 10.8 shows two dimensions along which influences on test scores can fall. To the degree that an influence on the test score is part of the construct we intend to measure, the influence increases the validity of the test. To the degree that an influence is stable over time, it increases the retest reliability of the test. Biases are stable influences that cause scores to be reliable but inaccurate. It is possible to have valid measurement of temporary states that, by definition, have low retest reliability. Traits are stable and valid influences on test scores. Error is unstable and not relevant to the construct.

The degree to which low reliability interferes with valid measurement is called *attenuation of validity*. Consider the tests in Figure 10.9. Test X_1 measures construct X, and test Y_1 measures construct Y. To make everything simple, all tests and constructs have a standard deviation of 1. The correlation between X and Y is ρ_{XY}. The relationship between the observed variables can be found by tracing the connecting paths between them, multiplying all the path coefficients as we go. Thus, the correlation between X_1 and Y_1 is:

$$\rho_{X_1Y_1} = \rho_{XY}\sqrt{\rho_{XX}\rho_{YY}}$$

Because reliability coefficients cannot exceed 1, the quantity $\sqrt{\rho_{XX}\rho_{YY}}$ cannot exceed 1, and therefore $\rho_{X_1Y_1} \leq \rho_{XY}$. To the degree that either measure produces unreliable scores, the observed correlation will likely be lower than its theoretical value.

If we have an estimate of reliability for both our measures and the estimated correlation between them, we can estimate how correlated the latent constructs are (Spearman, 1910):

$$\rho_{XY} = \frac{\rho_{X_1Y_1}}{\sqrt{\rho_{XX}\rho_{YY}}}$$

If this equation produces a value greater than 1, one or more estimates in the equation are inaccurate. When first learning about this method of disattenuating the observed validity coefficient, it seems impossible, and a little like cheating. However, this technique is essentially what occurs in any estimation of relationships between latent variables.

In psychological assessment, we often select tests that measure constructs that we know will explain (or predict) important life outcome variables. Each predictor's explanatory power depends on its correlation with the outcome measure. If we choose to measure predictors with low reliability, we have no hope of arriving at valid explanations or accurate predictions.

Understanding Composite Scores

Psychological tests and questionnaires can achieve adequate reliability only by adding many item scores together. Further reliability (and sometimes validity) is achieved by adding together scores from different tests to create *composite scores*. In all major cognitive and academic test batteries, various composite scores are created by default. Most—but not all—of these composite scores are created using sound logic and have solid evidential backing. To be confident that your test score interpretations are correct, you need a firm grasp of the principles by which good composites are created. Once this expertise is acquired, you can apply it by creating useful custom composite scores not anticipated by test publishers. If you are working within the Cattell-Horn-Carroll (CHC) framework (McGrew, 2009; Schneider & McGrew, 2018), detailed guidance for selecting tests for composite scores is found in the cross-battery assessment approach specified by Flanagan et al. (2013).

Good Composite Scores Are Theoretically Plausible

Not just any set of scores should be combined into a composite score. In most cases, composites scores are justified when the scores are strongly correlated and intended to be measures of the same well-validated construct. For example, the Oral Vocabulary test and the Picture Vocabulary test from the WJ IV (Schrank et al., 2014) both measure vocabulary. Although they measure different aspects of vocabulary via different test paradigms, combining them makes sense if our purpose is to measure a person's understanding of words rather than to emphasize the difference between the person's ability to define words orally and the person's ability to name pictured objects and concepts.

Good Broad Composites Have Adequate Content Validity

Most psychological constructs are too broad to be measured by a single item or even a single kind of test. *Content validity* refers to how adequately a measure samples the domain of the construct. For example, *working memory* refers to a domain-general capacity to hold and manipulate information in short-term memory (Baddeley, 2012). The construct has auditory, visual, and perhaps other kinds of components as well (e.g., episodic memory). If all your tests of working memory capacity are auditory, then your composite is a measure of auditory working memory, not working memory capacity in general. In many cases, that is exactly what you want, and in such cases, it is best to be clear about what your composite consists of.

Some constructs have so many components and facets that it is impractical to measure them all. In such cases, it is acceptable to select a few representative components, preferably those that are central to the construct's definition and have demonstrated utility. For example, Visual Spatial Processing (*Gv*) consists of many narrow abilities, most of which have unknown predictive validity. In such cases, it is best to select measures of visualization and mental rotation, which are core narrow abilities within *Gv* and have substantial research findings attesting to their utility.

Good Composites Are Well-Balanced

An imbalanced composite consists more of one kind of narrow component than another. If you have three tests in your broad ability composite, it is preferable to select tests measuring three different narrow abilities so that the composite is not imbalanced and has greater content validity. If you have two tests of one narrow ability and a single test of another narrow ability, you can maintain balance by creating a narrow composite with the first two tests and then combining the third test with the newly created narrow composite. With four tests, you can measure four different narrow abilities, or you can measure two narrow abilities with two tests each.

Sometimes composites are imbalanced for good reasons. For example, the Full Scale Intelligence Quotient (FSIQ) from the *Wechsler Intelligence Scale for Children, Fifth Edition* (WISC-V; Wechsler, 2015) consists of two measures of verbal comprehension (*Gc*), two measures of fluid reasoning (*Gf*), and one measure each of visual-spatial processing (*Gv*), working memory (*Gwm*), and processing speed (*Gs*). *Gc* and *Gf* are double-weighted because their regression coefficients are usually larger than that of other abilities when those scores are used to forecast a wide range of outcomes. Similarly, the WJ IV General Intellectual Ability (GIA) score differentially weights its component scores in proportion to their loadings on the general factor of intelligence. Thus, the score is more closely aligned with the construct it is intended to measure than it would be if the components were equally weighted.

If examiners are thinking of differentially weighting scores or including more scores of one narrow component than another, they need a compelling justification for doing so. For example, it is appropriate to include more inductive reasoning tests than general sequential reasoning tests in a fluid reasoning composite because inductive reasoning is more central to the fluid reasoning construct.

Good Composite Scores Consist of Tests with Diverse Paradigms

If your composite consists of tests using the same test paradigm, it will be harder to tell if a student's difficulty with the test is due to an ability deficit or because of a (possibly temporary) difficulty with the test paradigm. For example, the WISC-V Vocabulary and the Verbal Knowledge test from the *Stanford-Binet, Fifth Edition* (SB5; Roid, 2003) are both excellent but are nearly the same kind of vocabulary test. You might use one test as a follow-up for the other if the first test score is in doubt, but selecting both tests ahead of time as an overall measure of vocabulary would be a mistake.

Good Composites Exclude Measures of What the Composite Is Intended to Explain

We often use one construct to explain another, but the explanatory construct must not be a direct measure of what we hope to explain. For example, reading vocabulary tests have legitimate uses, but not for explaining word reading problems. To say that a low score on a reading vocabulary test *explains* the word reading deficit is to engage in circular reasoning. Reading vocabulary tests are better used as explanations of reading comprehension tests.

To take a less obvious example, suppose we believe that working memory deficits are interfering with a student's math calculation fluency. The *Paced Auditory Serial Addition Test* (PASAT; Gronwall & Sampson, 1974) affords a wonderful opportunity to observe in real time how a person attempts to perform mental arithmetic rapidly under a working memory load. However, because the PASAT requires rapid mental arithmetic, it should not be included in a working memory composite used to explain calculation difficulties.

For similar reasons, the language-heavy sentence repetition paradigm should not be included in a working memory composite that "explains" language problems. However, once a working memory deficit has been shown to be present with other tests, sentence repetition tests (e.g., WJ IV OL Sentence Repetition, *Wechsler Individual Achievement Test*, Fourth Edition (WIAT-IV; Wechsler, 2020) allow us to observe directly how working memory deficits interfere with receptive and expressive language.

Quantitative reasoning is considered a component of fluid reasoning (Carroll, 1993), but many quantitative reasoning measures blur the lines between cognitive and achievement measures because formal training in math confers obvious advantages on many test items. If a fluid reasoning composite is going to serve as an explanation of math difficulties, measures of quantitative reasoning should be separated from other fluid reasoning measures, particularly quantitative reasoning measures that require non-trivial calculations (e.g., WISC-V Arithmetic and WJ IV COG Number Series). Instead, use measures with no obvious calculation requirements to estimate general fluid reasoning. Overall, non-quantitative fluid reasoning deficits can serve as an explanation of quantitative reasoning deficits, which, in turn, explain difficulties with applied math problems.

Difference Scores

How unusual is it for a child's reading and math achievement scores to differ by two or more standard deviations? For example, suppose that a child's reading score is 85, and the child's math score is 115. How unusual is this pattern of scores? To answer this question precisely, we need a large sample of children comparable to this child. Access to such data sets is rare. However, it is unlikely that our answer needs to be precise. Approximate answers are usually sufficient for most assessment questions. We can use our knowledge of the properties of difference scores to obtain approximate answers to such questions. Suppose we have two variables, X (Reading) and Y (Mathematics). The difference score is:

$$D = X - Y = 85 - 115 = -30$$

When the X and Y have the same mean and standard deviation (σ), the mean of a difference score is 0, and the standard deviation of the difference score is:

$$\sigma_D = \sigma\sqrt{2 - 2\rho_{XY}}$$

Suppose that the correlation of X and Y is 0.60. Since both X and Y are standard scores with a standard deviation of 15, the standard deviation of the difference score becomes:

$$\sigma_D = 15\sqrt{2 - 2 \times 0.60} \approx 13.42$$

If both scores are approximately normal, the difference score is approximately normal. We can look up where a difference score is in the normal distribution by using the normal cumulative distribution function, which is available in most statistical programs and in Excel. In Excel:

=NORM.DIST(−30,0,13.42,TRUE)

=0.013

This probability means that about 1.3% of children differ by 30 or more points in this particular direction. If we double this amount to 2.6%, we estimate the proportion of children who differ by 30 or more points in either direction.

Difference Scores and Clinical Significance

Some test protocols and printouts will tell you if the difference between two scores is *statistically significant*. The significance value in this context tells you the probability that the scores would differ by the observed amount if their true scores were equal (i.e., the only reason the observed scores differ is due to measurement error). If the probability is low, then you can conclude that the two true scores differ by one or more points in the observed direction. With a term like *significance*, one would expect a more important finding than that. Indeed, statistical significance is relatively unimportant in the context of difference scores. Fairly small differences can be statistically significant if the scores' reliability coefficients are high. Small differences, even if statistically significant, are not likely to be clinically significant in terms of explanation, prediction, and treatment selection.

Because statistically significant differences are often quite common, it is better to focus on differences that are rare. Rare differences are not necessarily clinically significant, but they have the potential to be so. Common benchmarks for rarity are 15%, 10%, 5%, 1%, and powers of 10 smaller than that: 1 in 1000, 1 in 10,000, 1 in 100,000, and so forth.

The least controversial use of difference scores is that large (i.e., rare) differences should prompt us to check for scoring errors. The rarer the difference, the more likely some kind of measurement error has occurred. There are several other legitimate uses of difference scores. For example, some ability profiles are associated with choosing different kinds of college majors and careers. Scoring better in spatial ability and mathematics compared to verbal ability is associated with being more likely to choose a career in science, technology, engineering, and mathematics (Coyle et al., 2014).

It is important, however, not to get carried away with difference scores. Most differences are not meaningful in and of themselves, and small differences are unlikely to replicate in a subsequent evaluation (Styck et al., 2019). For this reason, elaborate hypotheses based on specific configurations of multiple scores are unlikely to be useful (McGill et al., 2018).

Some practitioners are taught that scores that are substantially different from each other should not be combined into composite scores. There is a measure of wisdom in this idea, but it is generally acceptable to combine discrepant scores as long as there is no reason to doubt the accuracy of the scores. A composite consisting of scores that are different from each other is generally just as valid as a composite consisting of scores that are similar to each other (Freberg et al., 2008; McGill, 2016; Schneider & Roman, 2018). The most likely reason that the scores differ from each other is that one score is a moderate overestimate of the ability, and the other is a moderate underestimate of the ability. More likely than not, the overestimate and the underestimate will cancel each other out, and the composite score will be closer to the true score than either test score alone. Also, the fact that two scores are similar to each other or nearly equal confers no protection against overestimation or underestimation. That is, it is quite possible that the two scores are both overestimates or both underestimates. If the scores are the same or different, the risk of inaccuracy is equal.

If you have reason to doubt one of the scores, then of course you should conduct follow-up tests with measures that assess similar abilities to see if one or both scores were inaccurate. If the score difference is confirmed with follow-up testing, it is legitimate to interpret the difference and, if needed, combine all available legitimate scores into a proper composite score.

The calculation method shown in this chapter compares scores one pair at a time. If you measure multiple abilities, there are many possible pairs of scores that might differ, and the probability that at least one pair differs by a large amount goes up when you measure more and more abilities. Although strongly correlated scores tend to be similar, uneven profiles are far more common than completely flat profiles. For example, if five abilities correlate at 0.60 (a common correlation size in the ability domain), about 94% of people will have at least one score difference of 10 or more. About 54% of people will have at least one pair of scores that differ by 20 or more, and 15% will have at least one pair of scores that differ by 30 or more.

Prediction

Sometimes psychologists are asked to forecast distal outcomes, such as the probability that a young child will eventually graduate from college. In such cases, most of us shy away from providing numerical estimates, even if we know how. Rarely do we have enough confidence in our causal models to take such estimates seriously. However, just as it is an error to be overconfident, it is a disservice to be coy when one does, in fact, have relevant information. Because we choose not to make a prediction does not mean that prediction errors will not be made. Decisions on a child's behalf are going to be made with or without our input. When called upon to make a prediction, we have an opportunity to help decision-makers think about prediction and uncertainty in more sophisticated ways.

Statistical prediction tools are appropriate to the degree that the underlying assumptions are consistent with one's conceptual causal model. Out-of-the-box prediction models such as regression often involve assumptions that are easy to overlook. We tend to use these tools mostly for back-of-the-envelope calculations because we rarely have enough confidence in our causal model that we

believe that the statistics will be highly accurate. Approximate answers that are treated as such can prevent professionals from making gross errors of interpretation. Of course, excessive confidence in one's statistics is equally dangerous, if not more so.

However, there is a kind of "prediction" that is really more a tool for explanation. When we use intelligence scores to help explain a child's academic performance, we are using, at least implicitly, a "prediction" equation. For example, if a child's IQ score is very low (e.g., 60), we "predict" that the child will also have poor academic skills. If the child's academic skills are indeed poor, we reason that the low IQ is a salient (but by no means only) factor in the explanation of the poor performance. If the child's academic skills are average despite low intelligence, the prediction has not "failed" but has actually revealed something important: something has gone especially right in the child's life. Now it is our job to understand what that something is.

Simple linear regression is a statistical procedure in which one continuous variable predicts another. Suppose we wish to forecast the likely range of IQ scores in a school-age child three years from now. The current score is 75. The exact stability coefficient depends on the test, ages, and population, but we can use the corrected stability coefficient of 0.84 from Watkins and Smith (2013) as a rough estimate. Over the short term, carryover effects need to be taken into account. However, in the Watkins and Smith (2013) study, the mean scores taken almost three years apart were essentially the same.

Calculating Predicted Values $\left(\hat{Y}\right)$

Although the notation is a bit different, the regression formula is the familiar formula for a line in an introductory algebra class:

$$\hat{Y} = b_0 + b_1 X$$

where:

$\hat{Y}$ = Predicted value of Y

b_0 = Intercept $\left(\hat{Y} \mid X = 0\right)$

b_1 = Slope

The equations for the coefficients are:

$$b_1 = \rho_{XY} \frac{\sigma_Y}{\sigma_X}$$

$$b_0 = \mu_Y - b_1 \mu_X$$

If both variables are z-scores (i.e., $\mu_X = \mu_Y = 0$ *and* $\sigma_X = \sigma_Y = 1$), $b_1 = \rho_{XY}$ and $b_0 = 0$, making the regression equation is a little easier to remember:

$$\hat{Z}_Y \Big| = \rho_{XY} z_X$$

That is, the predictor z-score times the correlation coefficient is the predicted z-score of the criterion variable. Memorizing this formula in combination with the z-score formula can come in handy for back-of-the-envelope estimations that can be performed anywhere. For example, suppose that $X = 85$, and X and Y correlate at 0.84. We can calculate the estimated score for Y like so:

$$z_X = \frac{X - u_X}{\sigma_X} = \frac{75 - 100}{15} = -\frac{5}{3}$$

$$\hat{Z}_Y = \rho_{XY} z_X = .84 \times -\frac{5}{3} = -1.4$$

$$\hat{Y} = \hat{Z}_Y \sigma_Y + \grave{\imath}_Y = -1.4 \times 15 + 100 = 79$$

Calculating the Standard Error of the Estimate (σ_E)

The fact that $\hat{Y} = 79$ is our point estimate is interesting, but it would be better if we had a sense of how precise the estimate is. The SEE is the standard deviation of the prediction errors:

$$\begin{aligned}\sigma_E &= \sigma_Y \sqrt{1 - \rho_{XY}^2} \\ &= 15\sqrt{1 - 0.84^2} \\ &= 8.14\end{aligned}$$

Conditional Distributions

Conditional distributions allow us to give approximate answers to questions such as, "What proportion of school-age children with an IQ of 75 will score at least in the average range (90 or better) when reevaluated three years later?" A *conditional distribution* is the distribution of one variable given certain conditions. In this case, we would like to know the distribution of *Y*, given a particular value of *X*. If $IQ_1 = 75$, the conditional distribution of IQ_2 is assumed to be normal, with a mean of $\widehat{IQ}_2 = 79$ and a standard deviation of $\sigma_e = 8.14$.

The normal cumulative distribution function will return the proportion of IQ_2 scores less than 90. Subtracting this quantity from 1 gives the approximate answer to the question: about 9% of children with an IQ of 75 will obtain an IQ of 90 or higher about three years later (see Figure 10.10).

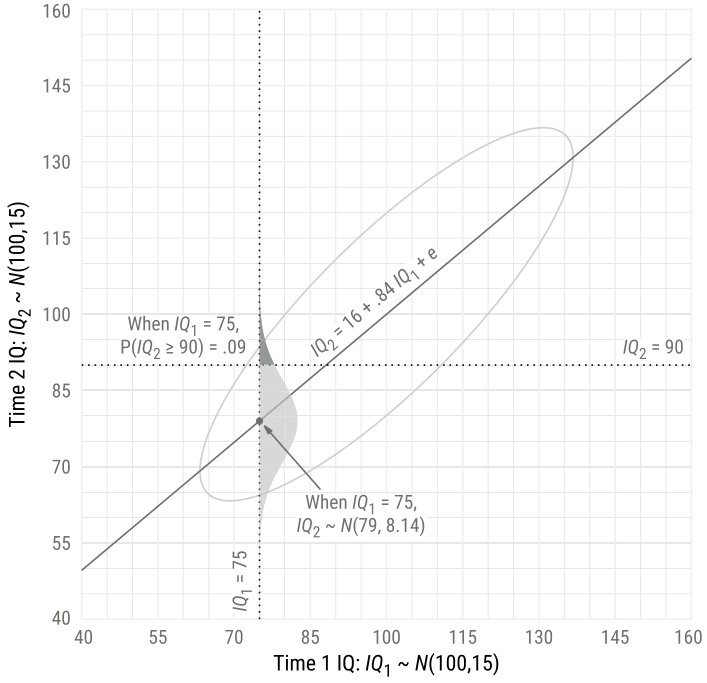

FIGURE 10.10 About 9% of children who score 75 on an IQ test will score 90 or better three years later.

SUMMARY

This chapter illuminates the underlying principles of measurement and psychometrics that are useful for the practitioner to know and understand. Guidance was provided in how foundational statistical concepts and analyses could be applied and utilized by practitioners conducting neuropsychological evaluations and completing interpretation via a neuropsychological interpretive and psychometric lens.

REFERENCES

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment: A psychological study of the strange situation*. Lawrence Erlbaum.

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29.

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs. *Intelligence*, *30*(6), 505–514. https://doi.org/10.1016/S0160-2896(02)00082-X

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295x.111.4.1061

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Burgaleta, M., Johnson, W., Waber, D. P., Colom, R., & Karama, S. (2014). Cognitive ability changes and dynamics of cortical thickness development in healthy children and adolescents. *NeuroImage*, *84*, 810–819. https://doi.org/10.1016/j.neuroimage.2013.09.038

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press. https://doi.org/10.1017/cbo9780511571312

Castellanos, F. X., Sonuga-Barke, E. J. S., Scheres, A., Di Martino, A., Hyde, C., & Walters, J. R. (2005). Varieties of attention-deficit/hyperactivity disorder-related intra-individual variability. *Biological Psychiatry*, *57*(11), 1416–1423. https://doi.org/10.1016/j.biopsych.2004.12.005

Charter, R. A., & Feldt, L. S. (2001). Confidence intervals for true scores: Is there a correct approach? *Journal of Psychoeducational Assessment*, *19*(4), 350–364. https://doi.org/10.1177/073428290101900404

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. L. Erlbaum Associates.

Coyle, T. R., Purcell, J. M., Snyder, A. C., & Richmond, M. C. (2014). Ability tilt on the SAT and ACT predicts specific abilities and college majors. *Intelligence*, *46*, 18–24.

Crawford, J. R., & Garthwaite, P. H. (2002). Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences. *Neuropsychologia*, *40*(8), 1196–1208. https://doi.org/10.1016/s0028-3932(01)00224-x

Crawford, J. R., Garthwaite, P. H., & Slick, D. J. (2009). On percentile norms in neuropsychology: Proposed reporting standards and methods for quantifying the uncertainty over the percentile ranks of test scores. *The Clinical Neuropsychologist*, *23*(7), 1173–1195. https://doi.org/10.1080/13854040902795018

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Cengage Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/bf02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). *Essentials of cross-battery assessment*. John Wiley & Sons.

Freberg, M. E., Vandiver, B. J., Watkins, M. W., & Canivez, G. L. (2008). Significant factor score variability and the validity of the WISC-III full scale IQ in predicting later academic achievement. *Applied Neuropsychology*, *15*(2), 131–139. https://doi.org/10.1080/09084280802084010

Furr, R. (2017). *Psychometrics: An introduction* (3rd ed.). SAGE.

Gronwall, D. M. A., & Sampson, H. D. (1974). *The psychological effects of concussion*. Auckland University Press.

Johnson, K. A., Kelly, S. P., Bellgrove, M. A., Barry, E., Cox, M., Gill, M., & Robertson, I. H. (2007). Response variability in attention deficit hyperactivity disorder: Evidence for neuropsychological heterogeneity. *Neuropsychologia*, *45*(4), 630–638. https://doi.org/10.1016/j.neuropsychologia.2006.03.034

Klein, C., Wendling, K., Huettner, P., Ruder, H., & Peper, M. (2006). Intra-subject variability in attention-deficit hyperactivity disorder. *Biological Psychiatry*, *60*(10), 1088–1097. https://doi.org/10.1016/j.biopsych.2006.04.003

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum. https://doi.org/10.4324/9781410601087

McGill, R. J. (2016). Invalidating the full scale IQ score in the presence of significant factor score variability: Clinical acumen or clinical illusion? *Archives of Assessment Psychology*, *6*(1), 49–79.

McGill, R. J., Dombrowski, S. C., & Canivez, G. L. (2018). Cognitive profile analysis in school psychology: History, issues, and continued concerns. *Journal of School Psychology*, *71*, 108–121. https://doi.org/10.1016/j.jsp.2018.10.007

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066x.50.9.741

Nunnally, J. C. (1967). *Psychometric theory*. McGraw-Hill.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge. https://doi.org/10.4324/9780203841624

Roid, G. (2003). *Stanford-binet intelligence test—fifth edition*. Riverside Publishing.

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll theory of cognitive abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (4th ed., pp. 73–130). Guilford Press.

Schneider, W. J., & Roman, Z. (2018). Fine-tuning cross-battery assessment procedures: After follow-up testing, use all valid scores, cohesive or not. *Journal of Psychoeducational Assessment*, *36*(1), 34–54. https://doi.org/10.1177/0734282917722861

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). *Woodcock-Johnson IV*. Riverside.

Spearman, C. E. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292. https://doi.org/10.2307/1412107

Spearman, C. E. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Styck, K. M., Beaujean, A. A., & Watkins, M. W. (2019). Profile reliability of cognitive ability subscores in a referred sample. *Archives of Scientific Psychology*, *7*(1), 119–128. https://doi.org/10.1037/arc0000064

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*(2), 174–195. https://doi.org/10.4135/9781412985789.n8

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2013). *Comprehensive test of phonological processing–2nd ed. (CTOPP-2).* PRO-ED.

Watkins, M. W., & Smith, L. G. (2013). Long-term stability of the Wechsler intelligence scale for children–fourth edition. *Psychological Assessment*, *25*(2), 477–483. https://doi.org/10.1037/a0031653

Wechsler, D. (2015). *Wechsler intelligence scale for children–fifth edition*. Pearson.

Wechsler, D. (2020). *Wechsler individual achievement test, fourth edition*. Pearson.