Chapter 9
Forensic Interviewing for Child Sexual Abuse: Why Psychometrics Matters

Scott O. Lilienfeld

The aim of science is not to open the door to everlasting wisdom, but to set a limit on everlasting error.

Bertolt Brecht (1955), Life of Galileo

It goes without saying that interviewers are human. As a consequence, they are susceptible to error. Nevertheless, because of *bias blind spot* (Pronin, Lin, & Ross, 2002), a phenomenon whereby most of us are keenly aware of biases in others but oblivious to the same biases in ourselves, interviewers may erroneously assume themselves to be largely immune to clinical mistakes that afflict their fellow interviewers. In light of research on bias blind spot, I offer the following bold proposition. The forensic interviewers who operate in the most scientific and ethical manner are *not* those who are free from error; instead, they are those who are cognizant of their propensities toward error and make tireless efforts to combat or compensate for them.

Errors that arise when interviewing children for potential child sexual abuse can be serious, even disastrous (Cronch, Viljoen, & Hansen, 2006). *False positive* errors, in which abuse is deemed to be present when it is absent, can contribute to unjustified allegations against parents and other caregivers; needless emotional suffering and stress for children and adults; inestimable damage to the reputations of innocent individuals; and wasteful expenditures of valuable legal, financial, and personal resources (Wood & Garven, 2000). *False negative* errors, in which abuse is deemed to be absent when it is present, can allow abusers to go free, and thereby increase the risk of abuse to other children. Of course, both types of errors can erode the credibility of the legal and mental health systems. For all of these reasons, it is imperative that psychologists, psychiatrists, social workers, and other would-be interviewers strive to minimize the risk of both types of error when conducting forensic assessments of child sexual abuse.

S.O. Lilienfeld, Ph.D. (✉)
Department of Psychology, Emory University, Room 473, Psychology and Interdisciplinary Sciences Building, 36 Eagle Row, Atlanta, GA 30322, USA
e-mail: slilien@emory.edu

W.T. O'Donohue, M. Fanetti (eds.), *Forensic Interviews Regarding Child Sexual Abuse*, DOI 10.1007/978-3-319-21097-1_9

Virtually all measures, including interviews, contain a certain degree of error. According to classical test theory, observed scores on a measure consist of two components: true score and error (Whiston, 2012). All things being equal, the lower the level of error, technically called *measurement error*, the more likely our instrument will be to detect the phenomenon of interest. Our fundamental goal as interviewers, both scientifically and ethically, should be to minimize error, and thereby maximize our odds of arriving a genuine picture of nature.

As the great American psychologist E. L. Thorndike (1918, p. 16) famously proclaimed, "Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality." Hence, Thorndike maintained, if we assume that a phenomenon exists, we can in principle measure it to some extent. For example, if we believe that a child is experiencing "underlying emotional conflicts" following sexual abuse, we should in principle be able to measure these conflicts, albeit imperfectly. To do so, we would first develop an *operationalization* of underlying emotional conflicts, followed by a measure of this operationalization (note that I used the term "operationalization" rather than the commonly used term, "operational definition," because the latter term implies erroneously that the operationalization is a strict dictionary-type definition of its intended construct; see Green, 1992). This measure would almost certainly be fallible, but it should hopefully be sufficiently saturated with the true score to be psychologically meaningful.

If a clinician were to aver that "I am certain that the child is experiencing underlying emotional conflicts, but there is no way to measure them," he or she would be in violation of Thorndike's dictum. Just as important, he or she would be operating unscientifically, because he or she would be advancing an assertion that is impossible to falsify (Popper, 1959). One major advantage of psychological measurement is that it forces us to be explicit in our assertions; more colloquially, it forces us to "put up or shut up" (see also O'Donoghue & Henderson, 1999). If our construct of "underlying emotional conflicts" is so nebulous and difficult to pin down that we cannot conceive of any way to operationalize, let alone measure, it, this should give us second thoughts regarding whether it is meaningful to begin with.

In this chapter, I review the basic principles of psychometrics, an applied technology that can help us to reduce, although not eliminate, the risk of error in clinical settings. I focus on classical test theory given its widespread use in forensic assessment; readers interested in generalizability theory, item response theory, and other more contemporary developments in psychometrics are referred elsewhere (e.g., Brennan, 2001; Embretson & Reise, 2000). Furthermore, I examine widespread sources of error in the forensic assessment of abuse and delineate constructive strategies for minimizing errors with the aid of psychological science.

The premise of this chapter is straightforward: Essentially all psychological measures are fallible, but we can minimize this fallibility by turning to psychometric methods, which are partial safeguards against errors. Furthermore, by relying on these safeguards, we can enhance the likelihood of accurate clinical decisions (Garb, 1998; Wood, Garb, & Nezworski, 2007). Paul Meehl (1997), the most influential clinical psychologist of the second half of the twentieth century, referred

to psychometrics as one of the few 'noble' tradiitions. in clinical psychology, largely because it is among the best established bodies of knowledge in our field (Wood et al., 2007). In this respect, it is sobering that the quality and quantity of psychometric training in most psychology departments, which is already grossly suboptimal, has stagnated or deteriorated in recent decades (Aiken, West, & Millsap, 2008; see also Borsboom, 2006). I hope that this chapter will play a modest role in reversing this trend.

The Bread and Butter of Psychometrics: Implications for Forensic Interviewing for Child Sexual Abuse

Psychometrics is the science of mental measurement. Sir Francis Galton, a cousin of Charles Darwin who is commonly regarded as the "father of psychometrics," was among the first to develop psychological tests to detect individual differences in intellect and personality. Galton was well aware of the problems posed by measurement error in individuals' observations of phenomena (Fancher, 2009). For example, Galton advocated the use of aggregation across multiple observers, which helps to cancel out random deviations in observations, to minimize error. In a classic demonstration in 1906, he asked 787 individuals at a livestock fair in England to estimate the weight of an ox on display. Although most individual estimates were wildly inaccurate, the mean of all estimates (1197 lb) was only 1 lb away from the ox's actual weight (1198 lb; see Surowiecki, 2005). In addition, recognizing that virtually all of us are poor at detecting covariation with our unaided eyes, Galton developed the technique of correlation, which his student Karl Pearson elaborated into a now-famous and widely used product–moment formula (Stigler, 1989). Galton's seminal contributions to mental testing and psychometrics fueled scholarly interest in the development of more sophisticated psychological measures, as well as of statistical techniques for evaluating them.

Because forensic interviews are interpersonal interactions that typically entail a substantial amount of subjectivity, it is easy to forget that they are first and foremost psychological measures. Hence, they are subject to the same psychological criteria as are all other measures. Virtually all students of psychology, as well as other mental health professionals, are well aware of the fundamental benchmarks, such as reliability and validity, needed to evaluate psychological measures, including forensic interviews. Nevertheless, many of these psychometric criteria are considerably more complex and nuanced than is commonly appreciated.

In the following section, I review fundamental psychometric principles relevant to forensic interviewing, with a particular emphasis on widely held misconceptions and misunderstandings that can impede scientifically grounded clinical assessment (see also Haynes, Smith, & Hunsley, 2011). Many standard psychometrics texts focus largely or exclusively on reliability and validity, but this duo is incomplete. We also need to consider the utility of psychological measurement (Haynes et al., 2011), often regarded as the undeservedly neglected stepchild of psychometrics.

Accordingly, I focus on the psychometric triad of reliability, validity, and utility. These three criteria are nested hierarchically: Reliability is necessary but not sufficient for validity, and validity is necessary but not sufficient for utility.

Reliability: Consistency of Measurement

In psychology, the term *reliability* refers to consistency of measurement. This term can generate all manner of confusion in forensic contexts, because in the legal system, "reliability" typically refers to the extent to which a measure is statistically associated with important outcomes, such as a deception, criminal risk, or childabuse (e.g., Leo, Drizin, Neufeld, Hall, & Vatner, 2006). Hence, in the courtroom, reliability generally means something more akin to what psychologists call validity.

Reliability: An Insider's Guide

The relation between reliability and validity is less straightforward than is typically assumed. Because reliability places constraints on validity, a measure that contains no reliable variance cannot be valid. Within classical test theory, scores on a measure that contains no reliable variance are composed entirely of random—that is, unsystematic—error (unsystematic errors are uncorrelated with each other). As a consequence, this measure cannot relate systematically to other variables.

Conversely, a measure that is extremely reliable can still be invalid for its intended purpose, as its reliable variance consists largely or entirely of error, namely systematic error (in contrast to unsystematic errors, systematic errors are intercorrelated). For example, imagine that we attempted to detect a history of child sexualabuse by inspecting children's figure drawings for signs of long, narrow objects, such as bullets or missiles; let us further imagine that children tended to be extremely consistent over time in whether they included such objects in their drawings. Our judgments of child sexual abuse would be highly reliable but entirely invalid, because there is no evidence that long, narrow objects are valid indicators of a history of sexual abuse. More generally, children's figure drawings are well-nigh useless for detecting sexual abuse (Lilienfeld, Wood, & Garb, 2000). Note, however, that I wrote "for its intended purpose" in the opening sentence of this paragraph. A measure that is highly reliable is almost certainly a valid measure of *something*, but it may not be a valid measure of the construct that the clinician or researcher has in mind (Sechrest, 1984).

Although higher reliability will, *all things being equal*, lead to higher validity, all things are not necessarily equal. Indeed, in some cases, increasing reliability can actually lower the validity of a measure, an important but little-known phenomenon called the *attenuation paradox* (Loevinger, 1954; see also Clark & Watson, 1995). This paradox can arise when a researcher attempts to boost a measure's reliability by

making it more homogeneous in content. For example, she might begin with a measure of childhood depression whose reliability (specifically, its internal consistency; see section "Internal Consistency") is deemed to be too low. To increase its reliability, she might write additional items designed to assess depressed mood and anhedonia (pervasive loss of pleasure), and jettison items designed to assess features of depression she regards as less central to the construct, such as concentration disturbances or psychomotor retardation (slowing). By doing so, she would probably end up with a more reliable but less valid measure of depression, because this measure would neglect to encompass the full range of signs and symptoms of this condition.

A further complexity is that validity is technically limited not by reliability per se but by the square root of reliability (Sechrest, 1984). Reliability can therefore technically exceed validity. For example, a psychiatric diagnosis can possess a reliability as low as $r = 0.6$ and in principle still display validities, as ascertained by correlations with other measures, as high as $r = 0.77$. This often overlooked point is potentially important, because some scholars have contended that as a field, we have sometimes overestimated the importance of reliability when evaluating psychological measures (Meehl, 1986). For example, the DSM-5 field trials almost exclusively emphasized reliability rather than validity (Regier et al., 2013), leaving open the possibility that a number of newly introduced diagnoses in the psychiatric manual are consistent but largely invalid measures of their intended constructs.

In reality, reliability and validity are not as distinct as we often imply. Instead, reliability and validity almost certainly lie on a continuum, namely, a dimension of *generalizability* (Campbell & Fiske, 1959; Cronbach, Rajaratnam, & Gleser, 1963). By generalizability, we mean the extent to which scores on the test can be extrapolated outside of the testing situation, such as to other measures, settings, or raters. At one extreme on the continuum, which comprises the "extreme" cases of reliability, one examines the association between maximally similar measures of the same construct. For example, across multiple adolescents, we might inquire about the levels of clinical depression twice within the same interview. Our resulting measure of association would be a prototypical measure of reliability, namely test–retest reliability (see section "Test–retest reliability"). At the other end of the continuum, the "extreme" cases of validity, one examines the association between maximally *dissimilar* measures of the same construct (Campbell & Fiske, 1959). For example, across multiple adolescents, we might inquire about their depression levels in an interview and also attempt to detect their levels of depression by administering an implicit association test. Our resulting measure of association would be an index of the validity of one or both measures. Between these two extremes, we often find cases that fall into the murky middle ground between reliability and validity. For example, if we attempt to detect adolescents' levels of clinical depression by administering an interview to them and then administering the same interview to their mothers (whom we ask to report on their child's depression levels), would the resulting index of association be an index of reliability or validity? From a scientific perspective, the answer is not especially important, because reliability and validity fall on a dimension of generalizability, with no clear line of demarcation between them.

Making matters still more complicated, measures can be consistent or inconsistent in different ways (Schmidt & Hunter, 1996). Hence, reliability is not a unitary concept (Sechrest, 1984). Moreover, for a given measure, different subtypes of reliability do not necessarily coincide in magnitude; levels of one form of reliability can be high while another is low. For example, scores derived from the Thematic Apperception Test, a widely used projective technique that asks respondents to tell stories in response to a series of ambiguous pictures, frequently display moderately high levels of test–retest reliability but low levels of internal consistency (Lundy, 1985). As a final complexity, we cannot assume that reliability values in one sample will necessarily generalize to other samples. For example, a self-report measure of attention-deficit hyperactivity disorder (ADHD) may display high reliability in an outpatient child disorders clinic, in which there is substantial variance in ADHD features, but low reliability in an undergraduate sample, in which the variance in ADHD features will presumably be lower (Haynes et al., 2011). As a consequence, reliability should be viewed as potentially conditional on our samples.

Hence, the commonly invoked, but hopelessly imprecise, phrase "this measure has been found to be reliable" should probably be forever banned from the pages of psychology journals, not to mention all courts of law. Authors and expert witnesses who are describing the reliability of their measures should instead be required to specify the form of reliability to which they are referring, as well as the sample from which it was derived.

Test–Retest Reliability

The best known form of reliability, *test–retest reliability*, refers to the stability of scores over time. In general, high levels of test–retest reliability for a measure are desirable. Nevertheless, this is only the case if one anticipates that the attribute being assessed should be reasonably stable over time, such as a personality trait (e.g., neuroticism) or a cognitive capacity (e.g., verbal ability). If one instead anticipates that the attribute in question should change over time, high levels of test–retest reliability would actually be undesirable. This distinction underscores a crucial point that holds for both test–retest reliability and the form of reliability I discuss next, namely, internal consistency: *The proper interpretation of reliability can only occur within a theoretical context*. That is, whether we expect or desire high levels of reliability hinges on our conceptualization of the construct being measured.

The question of the proper test–retest interval for a measure does not lend itself to a simple answer (Sechrest, 1984). Too brief a test–retest interval, such as one day, can be problematic, in part because individuals may recall their previous answers to questionnaires or interviews (Lord & Novick, 1968). Conversely, too lengthy a test–retest interval, such as several months, may also be problematic, as some of the changes in scores over time could reflect alterations in levels of true scores (the underlying attributes being measured, such as personality traits) rather than measurement error. Researchers and forensic assessors should be therefore certain to report the test–retest intervals used when presenting the test–retest reliability of a measure.

Internal Consistency

Measures of internal consistency address the question of how well the items on a measure "hang together" or, more precisely, assess the same construct. By definition, internal consistency applies only to measures that consist of multiple items. As in the case of test–retest reliability, high levels of internal consistency are generally desirable. Nevertheless, there are again exceptions. For example, a measure of independent life events, which are occurrences (e.g., death of a parent, experiencing an earthquake) that are presumably independent of the individual's behavior (Hammen, 1991), would be expected to display low internal consistency, because the events in question should be essentially random. Indeed, we might legitimately call into question an ostensible measure of independent life events that was highly internally consistent. Such high levels of internal consistency might suggest that the life events measure is contaminated by an unmeasured variable, such as negative emotionality (Watson & Clark, 1984).

One of the oldest measures of internal consistency is the *split-half coefficient*, which is calculated by dividing the test in half, and then correlating the two halves (Callender & Osburn, 1977). Most often, this procedure is accomplished by taking the odd numbered items on the test (1, 3, 5, etc.) and summing them, the even numbered items on the test (e.g., 2, 4, 6, etc.) and summing, and then correlating these two subtotals, yielding a statistic called *odd–even reliability*. Nevertheless, because there are many ways of divvying up a test into two halves, this procedure is somewhat arbitrary and can yield unstable results.

Hence, the preferred metric today for calculating internal consistency is Cronbach's alpha (Cronbach, 1951). Given some assumptions that we need not address here, Cronbach's alpha can be interpreted as the mean of all possible split-half coefficients (Cortina, 1993). Technically, we can interpret Cronbach's alpha within the context of parallel-forms reliability, the reliability reflecting the correlation between two parallel forms of the same test (Tavakol & Dennick, 2011). Specifically, if a measure has an alpha of 0.75, that value means that a measure consisting of the same number of psychometrically parallel items should correlate with the original measure at $r=0.75$. Cronbach's alpha is regarded as a "lower bound estimate" of internal consistency, because for multidimensional measures it will typically underestimate the interrelatedness among test items (Haynes et al., 2011).

Probably the most frequent error made in evaluating internal consistency estimates is to interpret them as indices of homogeneity (Cortina, 1993; Tavakol & Dennick, 2011). In fact, Cronbach's alpha is a notoriously poor indicator of homogeneity, especially for tests that contain many items. That is because Cronbach's alpha is affected substantially by test length. All things being equal, tests that contain more items will display higher internal consistencies as assessed by Cronbach's alpha and related metrics. Particularly for lengthy tests, authors should routinely report the mean inter-item correlation (MIC), which is simply the mean pairwise correlation among all items. In contrast to Cronbach's alpha, the MIC offers a direct index of homogeneity. For reasons that are unclear, MICs are rarely reported in published psychological research.

Inter-rater Reliability

The form of reliability most relevant to forensic interviews of child abuse is *inter-rater reliability*, the relation between the scores of two (or more) individuals, such as interviewers or other observers, on the same measure. Inter-rater reliability statistics address a fundamental question: Do the ratings of one interviewer generalize to those of other interviewers? If they do not, it raises the possibility that the scores of this interviewer are idiosyncratic.

When computing the inter-rater reliability of dimensional scores on interviews, such as interviewees' levels of anxiety or thought disorder, the preferred metric of agreement is the *intraclass correlation* (ICC). The ICC measures the amount of variance in scores attributable to the individuals being rated rather than to the raters themselves. If all of the variance in scores is attributable to the individuals being rated, the ICC will be 1.0; conversely, if all of the variance in scores is attributable to the raters, the OCC will be 0. In contrast to the Pearson product–moment correlation, which takes into account only the *relative* ranking and spacing of scores, most versions of the ICC are also influenced by the *absolute* levels of scores (McGraw & Wong, 1996). As a consequence, they will be influenced by differences in rater thresholds. For example, if two interviewers differ in their thresholds for labeling acts of physical aggression against children as "child abuse," with one requiring more severe or overt aggression than the other before labeling the behavior as abuse, the ICC, but not the Pearson correlation, will be affected by this difference.

When computing the inter-rater reliability of categorical scores on interviews, such as the presence or absence of child sexual abuse or the presence or absence of a psychiatric diagnosis (e.g., posttraumatic stress disorder), the preferred metric is the kappa coefficient, which measures the levels of chance-corrected agreement (Cohen, 1968; Fleiss & Cohen, 1973). By chance-corrected agreement, we mean agreement that is not attributable to the *base rates* (prevalences) of the phenomenon of interest (Brennan & Prediger, 1981). If we do not correct for chance-corrected agreement, we risk overestimating the level of inter-rater reliability. Imagine a case of two interviewers evaluating whether participants in an outpatient mood disorders clinic, in which the base rate of the diagnosis of major depression is 85 %, meet criteria for major depression. The interviewers could agree 85 % of the time merely by guessing that everyone in the sample meets criteria for major depression. By correcting for agreement that is potentially due to base rates, kappa addresses this problem. Nevertheless, because kappa may sometimes penalize raters for their shared expertise, it provides a statistically conservative estimate of rater agreement (Lilienfeld, Smith, & Watts, 2013).

Threats to the Reliability of Forensic Interviews

In a user-friendly and engaging analysis, Shea (1998) delineated a number of common threats to the reliability of interviews; most or all of these threats apply to

forensic interviews. I examine three such threats here. It should be borne in mind, however, that because reliability sets limits on validity, these errors also bear implications for the validity of interviews.

Cannon Questions

First, interviewers may engage in *cannon questions*, in which they "fire off" (Shea, 1998, p. 45) multiple queries in rapid succession, typically in the context of a single question (e.g., "Do you often feel extremely sad, tense, uptight, agitated, even suicidal?"). Such questions should almost always be avoided, because either positive or negative replies can be ambiguous or misleading. A "yes" reply could mean that the individual often experiences one of the emotions listed, some of them, or all of them; conversely, a "no" reply could mean that the individual denies experiencing one, some, or all of the emotions listed. Making matters worse, either a "yes" or "no" response could simply mean that the individual forgot some or all of the descriptors listed in the question. Cannon questions are especially ill-advised with children, who may misunderstand such questions, and with individuals with memory or concentration difficulties, including those with depression (see Watts, MacLeod, & Morris, 1988).

Phrasing Questions in the Negative

Second, phrasing questions in the negative (e.g., "You haven't thought about killing yourself, have you?"), which is a common error when inquiring about sensitive topics, can also diminish the reliability of interviews. Such questions, which again should almost always be avoided, can readily engender demand characteristics (Orne, 1962) in interviewees, as they can imply that the interviewer looks down on the behavior in question or is hoping for a denial of the undesirable behavior (Shea, 1998).

Altering Verbal or Nonverbal Behavior

Third, interviewers may subtly—or not so subtly—alter their verbal or nonverbal behaviors when asking certain questions during the interview (Shea, 1998). For example, when inquiring about sensitive topics, such as sexual abuse, physical abuse, drug use, or suicidal or homicidal ideation, they may lower their voice or change their pitch or speed of delivery of questions. Alternatively, they might inadvertently respond with signs of surprise, concern, or disapproval when the interviewee provides them an answer that is not to their liking. In all these cases, interviewers may unknowingly elicit inconsistent responses within respondents, across respondents, or both (Shea, 1998).

Validity: An Insider's Guide

Validity, as every psychology student learns, refers to the extent to which a measure assesses what it purports to measure; some students joke that the best way to answer a multiple choice question that concerns the definition for validity is simply to look for the word "purports." Recent years have witnessed the emergence of lively debates concerning the meaning and interpretation of validity (e.g., Borsboom, Mellenbergh, & van Heerden, 2004). I do not intend to revisit these at times arcane debates here, although I encourage interested readers to consult several recent discussions (e.g., Braun, 2012; Newton, 2012; Sechrest, 2005; Strauss & Smith, 2009).

The most crucial point for our purposes is that like reliability, validity is a multifaceted concept (Nunnally & Bernstein, 1994). Moreover, like reliability, validity is potentially conditional on the sample examined. Therefore, as in the case of reliability, the hackneyed phrase "this measure has been found to be valid" should be forever banished from psychological and psychiatric journal articles. I briefly review the major subtypes of validity here.

Content Validity

A measure's content validity refers to the extent to which it samples adequately from the "universe" of content comprising the construct of interest. For example, if I believe that psychopathic personality (psychopathy) is paradoxical constellation of characteristics that includes both psychologically adaptive features, such as superficial charm, interpersonal poise, and absence of anxiety, as well as psychologically maladaptive features, such as self-centeredness, guiltlessness, callousness, dishonesty, manipulativeness, and poor impulse control (see Cleckley, 1941; Hare, 1991/2003), a measure of psychopathy that consists only of maladaptive features would be of dubious content validity (see Lilienfeld, Patrick, Benning, Berg, Sellbom, & Edens, 2012, for a discussion). Although some authors have attempted to develop metrics to quantify content validity (see Polit, Beck, & Owen, 2007), these metrics have not caught on in most quarters. Hence, for better or worse, content validity is generally evaluated subjectively.

Content validity should not be confused with *face validity*, which is arguably not a form of validity at all (Lynn, 1986). Face validity refers to the extent to which test takers can infer the construct measured by the test. Face validity has long been a fraught concept in psychometrics, and for good reason. First, what may strike one test taker as obvious may strike another as obscure; hence, whereas one test taker may correctly surmise that the Beck Depression Inventory (Beck, Steer, & Brown, 1996) is a measure of clinical depression, another may assume that it is designed primarily to detect anxiety, suicide propensity, maladjustment, or negative emotionality. Second, in part for this reason, there is no standard quantitative metric for ascertaining face validity. Third, it is not even clear whether face validity is an

advantage or disadvantage for psychological measures. On the one hand, we might assume that *low* face validity would be advantageous, as this property should render it difficult for individuals to detect the purpose of the assessment and distort their responses accordingly (Bornstein, Rossner, Hill, & Stepanian, 1994). On the other hand, face validity is often associated with empirical validity. Indeed, efforts to generate self-report items with high empirical validity but low face validity, so-called *subtle items*, have typically been disappointing (Sechrest, 1984; Weed, Ben-Porath, & Butcher, 1990).

Criterion-Related Validity

Criterion-related validity is a broad concept that refers to the extent to which a measure relates to nontest variables (Maroof, 2012). The term "criterion-related validity" is typically preferable to the more traditional term "criterion validity" because there are precious few genuine "criteria"—infallible indicators or "gold standards"—in clinical psychology, personality, and allied fields (Cronbach & Meehl, 1955). For example, to ascertain the criterion-related validity of a measure of child sexual abuse, we might examine the extent to which it is associated with objectively corroborated indicators of abuse.

Criterion-related validity can itself be decomposed into several subtypes, corresponding to two overarching distinctions. First, we can subdivide criterion-related validity into subtypes corresponding to *when* the external variable was measured relative to the administration of the test. *Concurrent validity* examines whether the extent to which a test is associated with variables measured at about the same time the test was administered; *predictive validity* examines the extent to which a test is associated with variables measured long (e.g., months or years) after the test was administered; and *postdictive validity*, which is more rarely investigated, examines the extent to which a test is associated with variables measured long before the test was administered. For example, we might examine the concurrent validity of an interview-based measure of major depression by determining whether it correlates with a self-report measure of depression administered during the same session; we might examine its predictive validity by determining whether it correlates with future depressive episodes; and we might examine its postdictive validity by determining whether it correlates with past depressive episodes. In this case, both predictive and postdictive validity are premised on the fact that major depression tends to be an episodic and often recurrent disorder. Note that many authors misuse the term predictive validity, using it to refer to the extent to which a measure correlates with any nontest variable. This use is incorrect; this term should be reserved for the capacity of a measure to *forecast* future outcomes.

Criterion-related validity can be subdivided in another important way (Campbell & Fiske, 1959; Cole, 1987). *Convergent validity* examines whether a test correlates with measures of variables with which we would theoretically expect it to correlate. In contrast, *discriminant validity*, sometimes also called divergent validity, exam-

ines whether a test is uncorrelated (or largely uncorrelated) with measures of variables with which we would theoretically expect it not to correlate (or to correlate minimally). For example, if we were to develop a novel measure of posttraumatic anxiety for children, it would be important to demonstrate not only that the measure correlates positively with other measures of posttraumatic symptoms—convergent validity—but that the measure correlates less highly with variables that are theoretically unrelated or largely independent of such symptoms, such as intelligence or a social undesirability response style (see section "Threats to the validity of forensic interviews")—discriminant validity.

Given that most measures of psychopathology tend to be at least moderately positively correlated, tests of the discriminant validity of new measures of mental disorder are in many respects even more important than are tests of convergent validity (Tellegen, 1985). Virtually any measure of psychopathology will correlate at least moderately with other measures of psychopathology, even if it does not validly detect the construct of interest. For example, if I were to develop a new measure of depression that was actually more of a measure of anxiety, it would nonetheless correlate moderately with other measures of depression, because depression and anxiety measures are highly correlated (Dobson, 1985). As a result, I could be misled into concluding that my measure is a valid indicator of depression. If, however, I also administered a measure of anxiety, I would soon discover that my ostensible measure of depression correlated more highly with the anxiety measure than with another depression measure, revealing an absence of discriminant validity and forcing me to go back to the test construction drawing board. Nevertheless, discriminant validity tends to be underemphasized in the psychological literature.

Incidentally, many authors misuse the term discriminant validity to describe the capacity of a measure to discriminate between or among diagnostic groups. For example, many would describe the capacity of a measure of posttraumatic stress disorder (PTSD) to distinguish individuals diagnosed with PTSD from individuals diagnosed with another condition, such as major depression, as indicator of the measure's discriminant validity. In fact, it is an indicator of the test's convergent validity, because we are examining whether the measure correlates positively with another variable, namely, the presence versus absence of PTSD. The precise term for this psychometric property is *discriminative validity* (Haynes et al., 2011), which is a variant of convergent validity.

Construct Validity

Construct validity is the extent to which a measure detects a construct, which is a hypothesized attribute of individuals (Cronbach & Meehl, 1955; Loevinger, 1957). Constructs in clinical psychology include general intelligence, executive functioning, personality traits (e.g., extraversion), and psychiatric disorders (e.g., schizophrenia). None of these phenomena can be observed directly and can only be inferred. Because all of the forms of validity I have already reviewed bear on the

capacity of a measure to detect constructs, construct validity subsumes them. Hence, whenever we are measuring constructs—latent attributes—construct validity *is* validity (Messick, 1995; Sechrest, 1984; Waldman, Lilienfeld, & Lahey, 1995). Accordingly, authors who state that "this measure possesses good content, criterion-related, and construct validity" are asserting a pleonasm, not to mention committing a logical error. Construct validity supersedes these other forms of validity.

Construct validation requires test developers to postulate an explicit *nomological network*, a system of hypotheses that includes convergent and discriminant linkages among constructs, among variables, and between constructs and variables (Cronbach & Meehl, 1955; Waldman et al., 1995). For example, a researcher who developed a new measure of psychopathy should posit up front which variables he or she expects the measure to correlate with (e.g., current and future violence, diminished empathy as reported by self and others, psychophysiological indicators of fear insensitivity) as well as which variables he or she expects the measure to correlate weakly or at least less highly with (e.g., intelligence, depression, psychophysiological indicators of baseline arousal). The more evidence we amass over time that our measure correlates with theoretically predicted variables (convergent validity) and correlates weakly or not all with theoretically unpredicted variables (discriminant validity), the most compelling is the evidence for this measure's construct validity. Although provisional efforts have been made to quantify construct validity (Westen & Rosenthal, 2003), the evaluation of construct validity, like that of content validity, is almost always subjective.

Note that I wrote "explicit" in the first sentence of the previous paragraph. One of the hazards of construct validation, especially when it is performed in a less than rigorous manner, is that we can too easily accrue evidence for our measure in a post hoc fashion (Bechtoldt, 1959; Lynam & Miller, 2012). In other words, we can often "retrofit" evidence after the fact and claim that it was consistent with our initial hypotheses. Hence, it is incumbent on test developers to be as explicit as possible regarding which findings would falsify, or at least call into question, their assertion that their measure is a valid indicator of the intended construct.

Because construct validation, like the process of validating scientific theories, is in principle a continual and never-ending endeavor, we should avoid referring to measures as "validated." Instead, the best we can say is that extant evidence supports the assertion that our measure validly detects the latent attribute of interest.

Threats to the Validity of Forensic Interviews

A host of variables, some stemming from interviewers and others stemming from interviewees, can adversely affect the validity of interviews, including forensic interviews. Here I discuss three particularly important threats to interview validity. The first and third threats originate largely from interviewer behaviors, whereas the second threat originates largely from interviewee behavior. Nevertheless, because the interview is a dyadic interaction, all three errors can derive in part from the actions of both interviewer and interviewee.

Inadequate Probing

Inexperienced interviewers, and occasionally experienced interviewers who are experiencing intense time pressure, may commit the error of inadequate probing of interviewee responses. This mistake is especially likely when interviewers assume that they understand certain words or phrases on the part of interviewees, such as "depressed," "panicky," "aggressive," or "manic." As the American psychiatrist Harry Stack Sullivan (1954) noted in his classic book, *The Psychiatric Interview*, this assumption is almost always unwarranted, because these words or phrases do not necessarily have identical or even similar meanings across interviewees. One interviewee may say "I am feeling depressed" to refer to a mild state of sadness following a rough day at work, whereas another may use this phrase to refer to profound feelings of psychological agony. Similarly, one interviewee may report that "I was aggressive with my wife last night" to describe his interrupting her during an argument, whereas another may use this phrase to describe his physically assaulting her with a closed fist.

Perhaps the best antidote to this error is the use of *behavioral incidents* (Pascal, 1983; Shea, 1998), which are concrete behavioral examples or details. When eliciting behavioral incidents, interviewers probe interviewees' ambiguous terms and phrases by inquiring about specific actions. For example, rather than assuming that one understands what the interviewee means by "being aggressive," the skilled forensic interviewer would follow up with such probes as "In what ways were you aggressive?," "When you say 'aggressive', what do you mean?," "Tell me what you did," "What happened first?," "Then what happened?," and so on. Behavioral incidents can minimize the risk of error in interviews and thereby enhance their validity by enhancing the odds that interviewer judgments are grounded in reasonably objective behavioral indicators. Of course, the interviewees' selection of terms may itself sometimes be of clinical interest. An interviewee who habitually describes the physical abuse of his child as "being a bit rough with my kid every once in a while" may be engaging in minimization, a characteristic that may be tied to certain clinically important personality traits or personality disorders, such as psychopathy (Porter & Woodworth, 2007). Nevertheless, interviewers should not rely exclusively on the interviewers' choice of terms, as this reliance can be misleading.

Response Sets and Response Styles

Response sets and *response styles* are ways of responding to questions that are largely independent of content (Paulhus, 1991). Response sets and response styles fall on a continuum, with sets being primarily situational (e.g., a response to an insanity evaluation) and styles being primarily dispositional.

Response sets and styles, in turn, can be largely unsystematic or systematic (Piedmont, McCrae, Riemann, & Angleitner, 2000). Unsystematic response sets and styles, which are more relevant to self-report measures than to interviews,

include random or careless responding. Systematic response sets and styles, which are relevant to both self-report measures and interviews, include acquiescence/counteracquiescence, social desirability, and malingering. *Acquiescence*, colloquially called "yea-saying," reflects a propensity to answer yes to questions independent of their content; *counteracquiescence*, colloquially called "nay-saying," reflects a propensity to answer no to questions independent of their content. Acquiescence is a particular threat to the validity of forensic interviews with children, who are especially vulnerable to the effects of suggestive questioning (Bruck & Ceci, 2000; Ceci & Bruck, 1993). *Social desirability* is a propensity to provide answers that make oneself appear "good" in the eyes of others and to deny trivial faults (Ones, Viswesvaran, & Reiss, 1996). For example, a "yes" response to an item such as "I have no bad habits" would give a respondent a point on most standard social desirability scales. *Malingering* is virtually the opposite of social desirability and is a tendency to make oneself appear ill or psychologically disturbed. Many malingering scales consist of items designed to assess seemingly plausible features of psychopathology that are in fact exceedingly rare (e.g., Lilienfeld & Andrews, 1996). For example, a malingering item on a self-report scale might be "At times I see large fish, birds, or other animals floating in front of my eyes."

One often unappreciated advantage of self-report measures is that they can detect response sets and styles systematically, usually by means of embedded validity scales (Widiger & Frances, 1987). In keeping with the core theme of this chapter, the principle here is that *if one cannot eliminate a source oferror, one can at least attempt to measure it*. In turn, one can use systematic measures of response sets and response styles to compensate for error, such as by treating them analytically as moderators or suppressors of the validity of measures (see McGrath, Mitchell, Kim, & Hough, 2010, for a discussion). Interviews can in principle also be used to detect response sets and response styles, such as malingering, although relatively few systematic efforts have been undertaken in this regard (see Rogers, 2010, for a notable exception). Nevertheless, there is presently an active debate regarding whether controlling for social desirability and other response styles leads to clinically significant increases in net validity (McGrath et al., 2010; Piedmont et al., 2000; Rohling et al., 2011).

Suggestive or Leading Questions

Forensic interviewers can compromise validity by engaging in suggestive or leading questions, which fall on a dimension, with leading questions (e.g., "Daddy touched you there, right?") being more suggestive than suggestive questions (e.g., "I heard that Daddy touched you there. Is that right?"). Such questions can inadvertently end up providing interviewers with the answers they are looking or hoping for (Geiselman, Fisher, Cohen, & Holland, 1986). Nevertheless, these answers may be inaccurate. Suggestive or leading questioning can also contribute to low levels of inter-rater reliability, especially when some interviewers but not others engage in this practice. Ironically, a team of researchers or clinicians all trained to engage in suggestive or leading questioning could exhibit high levels of inter-rater reliability but low levels of validity.

Utility

Just as reliability does not ensure validity, validity does not ensure utility. Utility refers to the extent to which a measure is useful for clinical purposes. Utility addresses several important pragmatic questions, such as whether a measure enhances treatment outcomes (Hayes, Nelson, & Jarrett, 1987) or contributes to the statistical prediction of events, such as child abuse, above and beyond already collected data (Meehl, 1959). In the case of forensic interviews, an assessment of child sexualabuse should be clinically useful: It should help us to assign abused children to appropriate treatment, decrease their risk for subsequent psychopathology, and so on.

For reasons that are insufficiently appreciated, even a measure with extremely high validity may be virtually useless in certain clinical settings. This paradoxical state of affairs can arise in two major ways.

Base Rates

First, a measure that possesses high levels of criterion-related validity in one sample may be virtually clinically useless in a sample with an extremely low base rate of the phenomenon of interest (Meehl & Rosen, 1955; see also Finn & Kamphuis, 1995); as noted earlier, base rates refer to the prevalence of a phenomenon. Incidentally, it will also be virtually clinically useless in the rarer case of a sample with an extremely high base rate of this phenomenon. To take an extreme example, a measure with high validity for detecting sexual abuse will be clinically useless in a sample in which no one has been abused. Note, however, that this measure will similarly be clinically useless in a sample in which everyone has been abused. A measure cannot make differentiations if there is nothing to differentiate.

More commonly, of course, practitioners are tasked with the job of identifying a clinical phenomenon, such as sexual abuse, in a sample in which the base rate is not zero, but is very low. In such cases, a valid test may still yield little or virtually no clinically useful information. The mathematical formula known as *Bayes' theorem* reminds us that our proportion of correct identifications will be a joint function of (1) the test's validity and (2) the base rate of the phenomenon of interest (see Wood, 1996, for a superb tutorial on using Bayes theorem to inform child abuse evaluations). Although the mathematics of Bayes' theorem need not concern us here, suffice it to say that as base rates decrease, the rates of false positive identifications will increase. Moreover, if the base rates are sufficiently low, the use of a test with only modest validity can sometimes result in an *increase* in overall classification errors. In such situations, we would have been better off just "playing the base rates" and not using the test at all (Meehl & Rosen, 1955)!

Incremental Validity

Second, a valid measure may not be worth administering if it is redundant with other information, especially information that is already available to us. One of the

most crucial criteria for establishing clinical utility is *incremental validity* (Sechrest, 1963), the extent to which a measure is associated with clinically important outcomes above and beyond other measures. Given these considerations, it is surprising and perhaps disconcerting how rarely test developers attempt to demonstrate the incremental validity of newly constructed measures above and beyond extant measures (Garb, 2003; Hunsley & Meyer, 2003; Wood et al., 2007). For example, if an investigator were to develop a novel self-report measure for detecting child sexualabuse, the onus should be on him or her to demonstrate that this measure possesses "added value" above and beyond existing measures, especially measures that are less expensive and more easily administered. The lone major exception to this requirement is when new measures are designed to be briefer and more easily administered versions of existing measures with well-demonstrated validity. In such cases, a measure may not possess incremental validity above and beyond an extant measure; but if it is equally valid for detecting relevant phenomena, it should generally be preferred because it is more economical.

An important but rarely invoked distinction is that between *statistical* and *clinical* incremental validity. Statistical incremental validity refers to the extent to which a measure contributes additional statistical information, often quantified as a change in the amount of variance accounted for in a multiple regression equation, above and beyond extant information. Statistical incremental validity cannot be negative; at worse, it will zero. If a measure does not contribute additional statistical information above and beyond other measures, it will merely "drop out" of a regression equation, as all of its variance will have been soaked up by other measures.

In contrast, clinical incremental validity refers to the extent to which clinical judgments and predictions are enhanced by the addition of a new measure to an existing set of measures. Unlike statistical incremental validity, clinical incremental validity *can* be negative (Wedding & Faust, 1989). How? The literature on social cognition has identified a "dilution effect" whereby the provision of additional information sometimes results in an overall decrease in the accuracy of judgments (Nisbett, Zukier, & Lemley, 1981). This effect is especially likely to occur when the novel information is (1) more salient ("eye catching") than the existing information but (2) of lower validity than the existing information. Imagine that a forensic practitioner interested in ascertaining whether a client has been adversely affected by sexual abuse has collected a large body of psychometric data—biographical information, well-validated self-report measures of psychopathology, cognitive and other neuropsychological measures, observations from relatives and coworkers—and concluded that the evidence is inconclusive. Nevertheless, based on a brief, informal interview with the client that suggests maladjustment, the practitioner may be inclined to override the other data and place undue weight on the less systematic, and perhaps less valid, interview impressions. Indeed, the classic review of Sawyer (1966) suggested that the addition of interviews to additional psychometric information sometimes contributes to a net *decrease* in the accuracy of clinical judgments (see also Dana, Dawes, & Peterson, 2013).

This critical point is commonly misunderstood by individuals who reflexively recommend "more testing" whenever the answer to a clinical question (e.g., "What is the client's diagnosis?" "Was the client sexually abused?") is unclear. They may

assume that "more information is always better than less." Nevertheless, from the standpoint of clinical integration, this assumption is erroneous (Lilienfeld, Wood, & Garb, 2007). More information, especially if it is of lower validity than existing information, can inadvertently lead practitioners to rearrange their "mental regression weights" and accord unjustified emphasis to less valid data.

Norms and Standardization

Two other important, and closely related, means of reducing error in certain clinical inferences are the use of norms and standardization. Psychometricians commonly distinguish *criterion-referenced* from *norm-referenced* assessment (Popham & Husek, 1969). In criterion-referenced assessments, we are concerned only with *whether* a skill has been acquired or an attribute is present; we are not concerned with how the level of this skill or attribute compares with that of other individuals. A driver's test is a classic example of a criterion-referenced assessment. The governing body granting individuals a driver's license does not care how well a given driver performs relative to others drivers; all it cares about is whether the driver meets the established threshold for safe driving. Similarly, when assessing a child for a history of potential sexual abuse, the forensic interviewer is typically concerned only with whether this abuse is present.

In contrast, in norm-referenced assessment, we are preoccupied with the "compared with what" question (see Dawes, 1994). That is, we want to answer the question, "How do this person's scores compare with those of other people?" For example, when conducting a forensic interview, we may be interested in ascertaining how anxious a child is relative to other children of his or her age, and perhaps his or her gender. Or we may wish to determine whether a child who has been physically abused is experiencing more impaired executive functioning relative to nonabused children who have comparable overall levels of intelligence.

In such circumstances, accurate *norms* become crucial. Norms are average population baselines that form a basis of comparison with other scores (Cicchetti, 1994). Typically, norms are expressed in standard scores, such as scores that have a mean of 100 and a standard deviation of 15, as is the case with most standardized intelligence tests. In some cases, psychological tests may also be normed within specific subgroups, such as gender, race, or specific age subgroups.

The history of psychological assessment offers a powerful reminder of the need for accurate norms, as well as of the hazards of inaccurate norms. For example, the norms on some early intelligence tests were badly flawed, leading to numerous misclassifications of people with average of even above-average intelligence as intellectually disabled (Wood et al., 2007). Ironically, David Wecshler, the developer of the most widely used intelligence test used today, was classified as "feebleminded" by early intelligence tests. A more recent example comes from clinical practice and research on the Rorschach Inkblot Test. Data strongly suggest that the norms for the Comprehensive ("Exner") system, still the most widely used scoring and interpretative scheme for this test, are seriously in error and tend to misclassify

many psychologically healthy individuals as pathological (Wood, Nezworski, Garb, & Lilienfeld, 2001).

Standardization of administration, which increases the chances that measures are administered in a comparable fashion across respondents, is essential for accurate norms. For norms to be meaningful, we need to be certain that error variance relevant to methods of administration is minimized, and that the resulting scores on measures faithfully reflect individual differences in the construct of interest, such as intelligence.

In the case of forensic interviews for child abuse and other psychological phenomena, standardization per se is rarely relevant given that these interviews tend to be individually tailored to respondents. Nevertheless, forensic interviews, like other interviews, vary on a continuum from *unstructured* to *structured*. Unstructured interviews have few or no standard questions, probes, or algorithms (scoring criteria), whereas structured interviews fall on the opposite end of this dimension. Interviews that fall in between these two extremes are commonly called *semistructured*. In general, meta-analyses (mathematical syntheses of the literature) suggest that structured interviews possess higher inter-rater reliability and construct validity than do unstructured interviews (Schmidt & Zimmerman, 2004; Wiesner & Cronshaw, 1988), almost certainly because they reduce psychometric error arising from interviewer differences in (a) the initial questions and probe questions asked and (b) interpretation and scoring of answers.

Concluding Thoughts

The forensic interview, when well conducted, can yield remarkable amounts of clinically useful information. At the same time, the forensic interview is inevitably a fallible psychological instrument, conducted by fallible human beings. Fortunately, by attending carefully to psychometric principles, interviewers can reduce their risk of clinical errors and harmful outcomes, and hopefully arrive at a closer approximation of the state of nature. Psychometric principles help to keep us humble: They remind us of our propensities toward errors (see also McFall, 1997; O'Donohue & Lilienfeld, 2007). At the same time, these principles also steer us away from the abyss of nihilism, as they remind us that these errors can be partly remediated with the aid of the finely honed tools of clinical science.

References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63*, 32–50.

Bechtoldt, H. P. (1959). Construct validity: A critique. *American Psychologist, 14*, 619–629.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck depression inventory*. San Antonio, TX: The Psychological Corporation.

Bornstein, R. F., Rossner, S. C., Hill, E. L., & Stepanian, M. L. (1994). Face validity and fakability of objective and projective measures of dependency. *Journal of Personality Assessment, 63*, 363–386.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*, 425–440.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071.

Braun, H. (2012). Conceptions of validity: The private and the public. *Measurement: Interdisciplinary Research & Perspective, 10*(1–2), 46–49.

Brennan, R. L. (2001). *Statistics for social science and public policy: Generalizability theory*. New York, NY: Springer.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*, 687–699.

Bruck, M., & Ceci, S. J. (2000). The suggestibility of children's memory. *Annual Review of Psychology, 50*, 419–439.

Callender, J. C., & Osburn, H. G. (1977). A method for maximizing split-half reliability coefficients. *Educational and Psychological Measurement, 37*, 819–825.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–205.

Ceci, S. J., & Bruck, M. (1993). Suggestibility of the child witness: A historical review and synthesis. *Psychological Bulletin, 113*, 403–439.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309–319.

Cleckley, H. (1941). *The mask of sanity*. St. Louis, MO: Mosby.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology, 55*(4), 584–594.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory? *British Journal of Statistical Psychology, 16*, 137–163.

Cronch, L. E., Viljoen, J. L., & Hansen, D. J. (2006). *Forensic interviewing in child sexual abuse cases: Current techniques and future directions* (Faculty Publications, Department of Psychology, Paper 6). Lincoln, NE: University of Nebraska-Lincoln. Retrieved from http://digitalcommons.unl.edu/psychfacpub/6

Dana, J., Dawes, R. M., & Peterson, N. (2013). Belief in the unstructured interview: The persistence of an illusion. *Judgment and Decision Making, 8*, 512–520.

Dawes, R. D. (1994). *House of cards: Psychology and psychotherapy built on myth*. New York, NY: Free.

Dobson, K. S. (1985). The relationship between anxiety and depression. *Clinical Psychology Review, 5*, 307–324.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fancher, R. E. (2009). Scientific cousins: The relationship between Charles Darwin and Francis Galton. *American Psychologist, 64*, 84–92.

Finn, S. E., & Kamphuis, J. H. (1995). What a clinician needs to know about base rates. In J. N. Butcher (Ed.), *Clinical personality assessment: Practical approaches* (pp. 224–235). New York, NY: Oxford University Press.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613–619.

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association.

Garb, H. N. (2003). Incremental validity and the assessment of psychopathology in adults. *Psychological Assessment, 15*(4), 508–520.

Geiselman, R. E., Fisher, R. P., Cohen, G., & Holland, H. (1986). Eyewitness responses to leading and misleading questions under the cognitive interview. *Journal of Police Science and Administration, 14*, 31–39.

Green, C. D. (1992). Of immortal mythological beasts: Operationism in psychology. *Theory and Psychology, 2*, 291–320.

Hammen, C. (1991). Generation of stress in the course of unipolar depression. *Journal of Abnormal Psychology, 100*, 555–561.

Hare, R. D. (2003). *Manual for the Psychopathy Checklist-Revised (PCL-R)*. Toronto, Ontario: Multi-Health Systems (Original work published 1991).

Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1987). The treatment utility of assessment: A functional approach to evaluating assessment quality. *American Psychologist, 42*, 963–974.

Haynes, S. N., Smith, G. T., & Hunsley, J. D. (2011). *Scientific foundations of clinical assessment*. New York, NY: Routledge.

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15*, 446–453.

Leo, R. A., Drizin, S. A., Neufeld, P. J., Hall, B. R., & Vatner, A. (2006). Bringing reliability back in: False confessions and legal safeguards in the twenty-first century. *Wisconsin Law Review, 2006*(2), 479–538.

Lilienfeld, S. O., & Andrews, B. P. (1996). Development and preliminary validation of a self-report measure of psychopathic personality traits in noncriminal population. *Journal of Personality Assessment, 66*, 488–524.

Lilienfeld, S. O., Patrick, C. J., Benning, S. D., Berg, J. M., Sellbom, M., & Edens, J. F. (2012). The role of fearless dominance in psychopathy: Confusions, controversies, and clarifications. *Personality Disorders: Theory, Treatment, and Research, 3*, 327–333.

Lilienfeld, S. O., Smith, S. F., & Watts, A. L. (2013). Issues in diagnosis: Conceptual issues and controversies. In W. E. Craighead, D. J. Miklowitz, & L. W. Craighead (Eds.), *Psychopathology: History, diagnosis, and empirical foundations* (2nd ed., pp. 1–35). Hoboken, NJ: Wiley.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1*(2), 27–66.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2007). Why questionable psychological tests remain popular. *Scientific Review of Alternative Medicine, 10*, 6–15.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51*, 493–504.

Loevinger, J. (1957). Objective tests as instruments of psychological theory: Monograph supplement 9. *Psychological Reports, 3*, 635–694.

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

Lundy, A. (1985). The reliability of the thematic apperception test. *Journal of Personality Assessment, 49*, 141–145.

Lynam, D. R., & Miller, J. D. (2012). Fearless dominance and psychopathy: A response to Lilienfeld et al. *Personality Disorders: Theory, Research, and Treatment, 3*, 341–353.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*, 382–386.

Maroof, D. A. (2012). *Statistical methods in neuropsychology: Common procedures made comprehensible*. New York, NY: Springer.

McFall, R. M. (1997). Making psychology incorruptible. *Applied and Preventive Psychology, 5*(1), 9–15.

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin, 136*, 450–470.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46.

Meehl, P. E. (1959). Some ruminations on the validation of clinical procedures. *Canadian Journal of Psychology, 13*, 102–128.

Meehl, P. E. (1986). Diagnostic taxa as open concepts: Metatheoretical and statistical questions about reliability and construct validity in the grand strategy of nosological revision. In T. Millon & G. L. Klerman (Eds.), *Contemporary directions in psychopathology* (pp. 215–231). New York, NY: Guilford.

Meehl, P. E. (1997). Credentialed persons, credentialed knowledge. *Clinical Psychology: Science and Practice, 4*, 91–98.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194–216.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspective, 10*(1–2), 1–29.

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology, 13*, 248–277.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

O'Donohue, W., & Lilienfeld, S. O. (2007). The epistemological and ethical dimension of clinical science. In T. A. Treat, R. R. Bootzin, & T. B. Baker (Eds.), *Psychological clinical science: Papers in honor of Richard M. McFall* (pp. 29–52). New York, NY: Routledge.

O'Donohue, W., & Henderson, D. (1999). Epistemic and ethical duties in clinical decision-making. *Behaviour Change, 16*(1), 10–19.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*, 660–679.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*, 776–783.

Pascal, G. R. (1983). *The practical art of diagnostic interviewing*. Homewood, IL: Dow Jones-Irwin.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson & P. R. Shaver (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic.

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology, 78*, 582–593.

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health, 30*, 459–467.

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*, 1–9.

Popper, K. R. (1959). *The logic of scientific discovery*. New York, NY: Basic Books.

Porter, S., & Woodworth, M. (2007). "I'm sorry I did it… but he started it": A comparison of the official and self-reported homicide descriptions of psychopaths and non-psychopaths. *Law and Human Behavior, 31*, 91–107.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28*, 369–381.

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2013). DSM-5 field trials in the United States and Canada, part II: test-retest reliability of selected categorical diagnoses. *American Journal of Psychiatry*, *170*, 59–70.

Rogers, R. (2010). *Structured interview of reported symptoms*. New York, NY: Wiley.

Rohling, M. L., Larrabee, G. J., Greiffenstein, M. F., Ben-Porath, Y. S., Lees-Haley, P., Green, P., & Greve, K. W. (2011). A misleading review of response bias: comment on McGrath, Mitchell, Kim, and Hough (2010). *Psychological Bulletin*, *137*, 708–712.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66*, 178–200.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199–223.

Schmidt, F. L., & Zimmerman, R. D. (2004). A counterintuitive hypothesis about employment interview validity and some supporting evidence. *Journal of Applied Psychology, 89*, 553–561.

Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement, 23*, 153–158.

Sechrest, L. (1984). Reliability and validity. In I. A. S. Bellack & M. Hersen (Eds.), *Research methods in clinical psychology* (pp. 24–54). New York, NY: Pergamon.

Sechrest, L. (2005). Validity of measures is no simple matter. *Health Services Research, 40*, 1584–1604.

Shea, S. (1998). *Psychiatric interviewing: The art of understanding*. Baltimore, MD: Saunders.

Stigler, S. M. (1989). Francis Galton's account of the invention of correlation. *Statistical Science, 4*, 73–79.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*, 1–25.

Sullivan, H. S. (1954). *The psychiatric interview*. New York, NY: W.W. Norton & Company.

Surowiecki, J. (2005). *The wisdom of crowds*. New York, NY: Random House.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55.

Tellegen, A. (1985). Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In A. H. Tuma & J. D. Maser (Eds.), *Anxiety and the anxiety disorders* (pp. 681–706). Hillsdale, NJ: Erlbaum.

Thorndike, E. L. (1918). Fundamental theorems in judging men. *Journal of Applied Psychology, 2*, 67–76.

Waldman, I. D., Lilienfeld, S. O., & Lahey, B. B. (1995). Toward construct validity in the childhood disruptive behavior disorders: Classification and diagnosis in DSM-IV and beyond. In T. H. Ollendick & R. J. Prinz (Eds.), *Advances in clinical child psychology* (Vol. 17, pp. 323–364). New York, NY: Plenum.

Watson, D., & Clark, L. A. (1984). Negative affectivity: the disposition to experience aversive emotional states. *Psychological Bulletin, 96*, 465–490.

Watts, F. N., MacLeod, A. K., & Morris, L. (1988). Associations between phenomenal and objective aspects of concentration problems in depressed patients. *British Journal of Psychology, 79*, 241–250.

Wedding, D., & Faust, D. (1989). Clinical judgment and decision making in neuropsychology. *Archives of Clinical Neuropsychology, 43*, 233–265.

Weed, N. C., Ben-Porath, Y. S., & Butcher, J. N. (1990). Failure of Wiener and Harmon Minnesota Multiphasic Personality Inventory (MMPI) subtle scales as personality descriptors and as validity indicators. *Psychological Assessment, 2*, 281–285.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology, 84*, 606–618.

Whiston, S. (2012). *Principles and applications of assessment in counseling*. Belmont, CA: Cengage Learning.

Widiger, T. A., & Frances, A. (1987). Interviews and inventories for the measurement of personality disorders. *Clinical Psychology Review, 7*, 49–75.

Wiesner, W. H., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61*, 275–290.

Wood, J. M. (1996). Weighing evidence in child sexual abuse evaluations: An introduction to Bayes' theorem. *Child Maltreatment, 1*, 25–36.

Wood, J. M., Garb, H. N., & Nezworski, M. T. (2007). Psychometrics: Better measurement makes better clinicians. In S. O. Lilienfeld & W. T. O'Donohue (Eds.), *The great ideas of clinical science: 17 principles that every mental health professional should understand* (pp. 77–92). New York, NY: Routledge.

Wood, J. M., & Garven, S. (2000). How sexual abuse interviews go astray: Implications for prosecutors, police, and child protection services. *Child Maltreatment, 5*, 109–118.

Wood, J. M., Nezworski, M. T., Garb, H. N., & Lilienfeld, S. O. (2001). The misperception of psychopathology: Problems with the norms of the comprehensive system for the Rorschach. *Clinical Psychology: Science and Practice, 8*, 350–373.