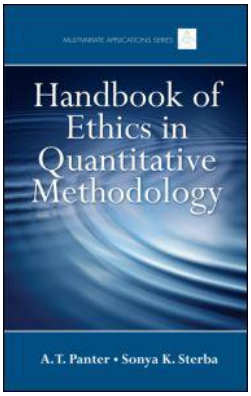

Handbook of Ethics in Quantitative Methodology

A.T. Panter, Sonya K. Sterba

Measurement Choices: Reliability, Validity, and Generalizability

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9780203840023.ch5
Madeline M. Carrig, Rick H. Hoyle
Published online on: 20 Jan 2011

Section III

Ethics and Research Design Issues

5

Measurement Choices: Reliability, Validity, and Generalizability

Madeline M. Carrig
Duke University

Rick H. Hoyle
Duke University

The choice of measurement instrument is a critical component of any research undertaking in the behavioral sciences and is a topic that has spawned theoretical development and debate virtually since the dawn of our field. Unlike the eminently observable subjects of many other fields of scientific inquiry—for example, the physical characteristics of rock cores in sedimentary stratigraphy or the velocity of blood flows in biomedical engineering—the subject of interest in behavioral research is often human thoughts, feelings, preferences, or cognitive abilities that are not readily apparent to the investigator, and which may even be out of the full awareness of the research participant. Over the years, many hundreds of tools, such as pencil-and-paper questionnaires, projective tests, neuropsychological batteries, and, more recently, electrophysiological and neuroimaging techniques, have been developed or tailored in an attempt to capture the essence of various behavioral phenomena. For the research (or indeed, applied) behavioral scientist, the question arises: When it is time to operationalize a behavioral construct of interest, how should I choose and implement an instrument in a way that is consistent with ethical practice?

Practitioners often look to their governing associations for guidance on matters of professional ethics, and fortunately, in its 2002 *Ethical Principles of Psychologists and Code of Conduct* (the ethics code), the American Psychological Association (APA) provides some beginning guidance in answer to this question. In the sections of the ethics code that are most relevant to the ethical selection and use of behavioral measurement instruments in research, the code states:

1. Psychologists administer, adapt, score, interpret, or use assessment techniques, interviews, tests, or instruments in a manner and for purposes that are appropriate in light of the research on or evidence of the usefulness and proper application of the techniques (Section 9.02.a, p. 13).
2. Psychologists use assessment instruments whose validity and reliability have been established for use with members of the population tested. When such validity or reliability has not been established, psychologists describe the strengths and limitations of test results and interpretation (Section 9.02.b, p. 13).

Hence, we are reminded that it is ethical to select instruments that are *useful* and *properly applied*. We are particularly encouraged to administer measures whose *reliability* and *validity* have been established in the population of interest and to report supporting psychometric evidence. But what types of evidence are most germane? More fundamentally, how should the research behavioral scientist evaluate whether an instrument, as well as his or her application of that instrument, possesses the desired characteristics?

It is to this latter question that the present chapter is substantially devoted. Our overarching goal is to provide guidance to the research behavioral scientist on the ethical selection and implementation of behavioral measurement instruments. We begin with a discussion of reliability and validity—two properties of measurement instruments that promote usefulness and proper application—and provide an overview of the methods that are presently available to the research behavioral scientist for the assessment of these properties. With respect to the proper application of such instruments, we also address the importance of considering the level of measurement, especially when instruments are involved in quantitative data analysis. Next, we expand our discussion of the ethics of behavioral measurement in research to include a survey of current scientific and ethical reporting standards. We then provide a summary of recommendations for practice. Finally, we present a case example, with the aim of highlighting key features of ethical conduct.[1]

[1] Other issues pertinent to the issue of the ethics of behavioral measurement pertain more specifically to psychodiagnostic assessment as performed by the clinical or school psychologist, such as training and supervision issues, use of tests for diagnosis, and the security of test materials and results. Such issues are addressed by sections of the APA code not presented here and are also discussed in detail in Koocher and Keith-Spiegel's excellent 2008 text. See also Wright and Wright (2002) for an interesting discussion of the ethics of behavioral measurement that focuses on the participant as a research stakeholder.

Reliability and Validity

As is reflected in the APA ethics code, it is generally agreed that the two most desirable properties of a behavioral measurement instrument are that instrument's reliability and validity. The original developer of a behavioral measurement instrument bears a responsibility for furnishing reliability and validity evidence that supports the use of the instrument for its stated purpose, and it is reasonable for the investigator to consider that evidence when making a selection among instruments. However, the investigator ultimately bears the responsibility of demonstrating the reliability and validity of the instrument in the particular setting in which it has been used (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999). Correspondingly, most recent conceptualizations of these desirable psychometric properties focus more strongly on the reliability and validity of a particular measure's implementation, rather than on assessment of the validity and reliability of the measure per se, as will be highlighted below.

Reliability

Reliability may be defined as the consistency of measurement instrument scores across replications of the measurement procedure (Brennan, 2001). Fortunately, it is a property of measurement that lends itself directly to quantification and statistical evaluation. Perhaps less fortunately, a dizzying array of methods for quantifying reliability are available, most of which depend on adoption of particular statistical models of measurement and/or definitions of the set of replications across which reliability will be assessed. Most of these methods involve the computation of either a *standard error of measurement* or the estimation of a *reliability coefficient*. We provide a brief overview of the various approaches. Our discussion draws on the comprehensive chapter written by Haertel (2006), which itself draws on earlier works by Thorndike (1951), Stanley (1971), and Feldt and Brennan (1989).

Classical Test Theory

In classical test theory (CTT), the model $X = T + E$ is used to describe the relationship between an observed score X, a "true" (error-free) score T, and the total measurement error E, where E may arise from any number of sources but is assumed to be uncorrelated with the true score T (Lord & Novick, 1968). In CTT, the reliability coefficient may be defined as the proportion of the total variance in observed scores that can be attributed to

true-score variance, or equivalently, as the squared correlation between the observed and true scores. As such, the reliability coefficient will assume values between 0 and 1 inclusive, with larger values reflective of greater reliability.

Estimation of the Reliability Coefficient

Although the reliability coefficient of a particular measurement process is rarely—if ever—exactly known, it may be numerically estimated. Over the years, CTT has given rise to multiple methods for producing such estimates. These methods, reviewed by Haertel (2006) in detail, include (a) the *parallel forms* reliability estimate, which is the correlation of scores resulting from two interchangeable (parallel) forms of a single measurement instrument administered to a single sample of participants at two points in time; (b) the *test–retest* reliability estimate, which is the correlation of scores resulting from two identical forms of a single measurement instrument administered to a single sample of participants at two points in time; and (c) the *staggered equivalent split-half procedure* (Becker, 2000), which attempts to take advantage of parallel-forms reliability estimation under circumstances when only one form of the measurement instrument is available.

An especially large category of methods for estimating the reliability coefficient in CTT includes *internal consistency* estimates, which tend to be frequently used because they were developed for the assessment of reliability from a single administration of a measurement instrument. All forms of internal consistency estimation involve subdividing the items of a measurement instrument and then observing the consistency of scores across subdivisions. Types of internal consistency estimates include (a) estimates that rely on the subdivision of the instrument into two parts, such as the *Spearman–Brown*, *Flanagan* or *Guttman–Rulon split-half*, *Raju*, and *Angoff–Feldt* coefficients (with Feldt & Charter, 2003, providing some guidance on making the best selection between them); and (b) estimates that rely on the subdivision of the measurement instrument into more than two parts, including *coefficient alpha*, *Kuder–Richardson 20*, *Kuder–Richardson 21*, *standardized alpha*, and *Guttman's* λ_2. Although popular, internal consistency estimates are likely to overestimate a measurement instrument's reliability because they do not capture error associated with possible fluctuations over time in responses to the instrument. The reader is encouraged to consult Haertel (2006) for references and for technical and computational details. Haertel (2006) also addresses estimates of reliability that are appropriate for composite scores, including difference scores, and provides information on computation of the CTT *conditional standard error of measurement*, which provides the standard error of measurement for a particular true score and is therefore useful for computing true-score confidence intervals.

Applications of Reliability Estimation in Statistical Analysis

When we apply inferential statistical models, we are generally interested in investigating relationships among the true scores on the constructs we intended to measure. However, models are generally fit to observed scores, and because of the complexities of assessing intra- or extrapsychic human behavior, even the best-conceived behavioral measurement instrument is likely to fail to achieve perfect reliability. Unfortunately, use of observed scores that are not perfectly reliable in the context of inferential statistical models can produce seriously misleading results, with potentially dramatic repercussions on the development of theory, clinical practice, policy, and the direction of future research. Failure to account for the presence of measurement error in a covariate used within an analysis of covariance (ANCOVA) model, for example, can lead either to significant *F* tests in the presence of no true adjusted effect or to nonsignificant *F* tests in the presence of a true adjusted effect (Maxwell & Delaney, 2004). Cohen, Cohen, West, and Aiken (2003) point out the potential of instrument fallibility to distort partialled relationships (e.g., partial regression coefficients) and to increase Type I or Type II error rates in the more general multiple regression/correlation analysis framework. Likewise, via simulation results, Hoyle and Kenny (1999) have demonstrated that mediational analyses that fail to account for unreliability in the mediating variable can produce biased parameter estimates and increase Type I and Type II error rates for the associated statistical tests.

The real threat of unreliability to the correctness of statistical conclusions under many circumstances has led to the development of statistical frameworks within the CTT tradition that attempt to "correct" for observed scores' fallibility, providing measures of effect that more closely reflect the relationships among the true scores (constructs) under investigation. Huitema (1980), for example, addresses options for analysis that may correct the problem in the context of ANCOVA; Cohen et al. (2003) detail methods developed to correct for the attenuation of correlation coefficients associated with measurement error and provide an overview of the strengths and weaknesses of existing remedies for the distortion of partialled relationships. Furthermore, Hedges and Olkin (1985) address correction for unreliability-associated attenuation of effect size.

The methods just described apply after-the-fact adjustments to parameter estimates produced from fitting a statistical model to a set of fallible observed scores. In general, they rely on the assumption that a measurement instrument's reliability is known. However, the substitution of estimated reliabilities can lead to potentially problematic results (see, e.g., Dunivant, 1981). Fortunately, more sophisticated methods are available that account for measurement error during the estimation process itself. Structural equation modeling procedures (Jöreskög, 1970), for example,

allow for the specification of a measurement model in which an unobserved *latent variable*, which holds an individual's hypothetical true score, and a separate unobserved measurement error variable together predict the individual's observed score on each of a number of behavioral measures. Relationships among the unobserved latent variables—the true-score measures of the constructs of interest—may then be modeled as the investigator sees fit, with the resulting parameter estimates presumably being free from the deleterious effects of measurement error. *Instrumental variable* estimation may also be used to minimize, or perhaps even remove, the negative influence of measurement error on parameter estimates (e.g., Hägglund, 1982).

Generalizability Theory

Because of the relative complexity of its associated models and data-analytic methods, *generalizability theory* (GT) will perhaps be less familiar to the research behavioral scientist than CTT. GT is largely (although not universally) viewed as an extension of CTT. Haertel (2006) provides a brief but readable introduction, and Brennan (2001) offers a more comprehensive treatment.

As noted above, the basic CTT measurement model includes one term (E) that captures the total of measurement error. Relative to CTT, GT offers many advantages in terms of the evaluation of a measurement process's reliability. Perhaps the two most important are (a) the inclusion in measurement models of terms that permit the specification of multiple and distinct types of error and (b) a more precise conceptualization of the set of replications across which reliability is to be evaluated.

Even a very basic application of GT to a reliability evaluation requires multiple definitions and decisions. For example, the investigator must identify the potential sources of error variance in the observed scores. These might include, for example, rater, test form, location of administration, and occasion of measurement. In GT, each source (e.g., rater) is named a *facet*, and each level within that source (e.g., Jane, Bill) is considered a *condition* of that facet. The investigator must also specify a so-called *universe of generalization*, defining the exact set of potential replications across which reliability will be defined for a particular measurement process. Accordingly, a single "measurement" within the universe of generalization might include a set of multiple observations (i.e., a collection of observed scores), each associated with a particular condition for each facet. Importantly, the investigator must also decide whether each of his or her facets is *random* or *fixed*. Random facets are those for which the particular conditions observed by the investigator in one measurement are viewed as a random sample from an infinitely large population of conditions to which the investigator seeks to generalize. Fixed facets, on the other hand,

are those involving a set of conditions that will not vary across the set of hypothetical measurements within the universe of generalization.

In a GT *decision study (D-study)*, a basic linear measurement model might explain an observed score as a function of a person (participant) effect, multiple facet effects, and perhaps effects that represent interactions among effects (together with a residual). Random-effects analysis of variance (ANOVA) is used to estimate the *variance components* (variances) of the various effects included in the measurement model. This procedure permits the estimation of a *universe score variance*, which captures the variability of the person effect across the hypothetical measurements in the universe of generalization. Estimated variance components can be used to compute *coefficients of generalizability*, which assess the reliability of a particular measurement instrument within the defined universe of generalization. Under some circumstances, certain generalizability coefficients (e.g., the $E\rho^2$ of Cronbach, Gleser, Nanda, & Rajaratnam, 1972) simplify to forms of the reliability coefficient defined in CTT. Extensions of GT for more complicated measurement models and data structures are available (cf. Brennan, 2001). Haertel (2006) provides a brief overview of the estimation of conditional standard errors of measurement from the perspective of GT.

Item Response Theory

The *item response theory* (IRT) model (e.g., Lord, 1968; Lord & Novick, 1968)—sometimes also named the latent trait model (Lord, 1953), logistic test model (Birnbaum, 1968), or Rasch model (Rasch, 1960)—is a family of models that uses a function of a set of participant and item parameters to describe the probability that a participant will receive a particular score on an individual measurement instrument item. Specific models within the IRT family may be differentiated in terms of multiple characteristics, including (a) the type of score produced by the measurement instrument items (i.e., binary vs. a *polytomous*, or ordered-categorical, outcome); (b) the model's *dimensionality*, or in other words, the number of participant parameters (also known as *abilities*, *traits*, or *proficiencies*) included in the model; (c) the number and type of item parameters involved in the model (which may include, e.g., characteristics such as item difficulty or capacity to discriminate among participants of differing abilities); and (d) the particular mathematical function used to relate the participant and item parameters to the observed score (Yen & Fitzpatrick, 2006).

The IRT model may be distinguished from the CTT and GT models in multiple ways, including (a) the IRT model's greater focus on item versus test-level scores; (b) the IRT model's somewhat more restrictive definition of a replication, with all item parameters in the IRT framework typically viewed as being fixed across all possible replications; (c) differences across models in the exact meaning of "true score"; and (d) the lack of an error

term in IRT (cf. Brennan, 2006). Hambleton and Jones (1993) note that the assumptions made by the IRT model are relatively more difficult to satisfy than those of CTT but emphasize that if the model fits the observed data well, IRT offers the advantage of participant and item parameters that are sample independent. Brennan summarizes his view of the differences between IRT, CTT, and GT thusly: "IRT is essentially a scaling model, whereas classical test theory and generalizability theory are measurement models. The essential difference, as I see it, is that a measurement model has a built-in, explicit consideration of error" (p. 6).

In great part because of its model's lack of an error term, IRT does not provide the more traditional reliability coefficients offered by CTT and GT. However, the IRT test information function does yield its own version of the conditional standard error of measurement, with technical and computational details addressed by Yen and Fitzpatrick (2006). Although the IRT conditional standard error of measurement is often used in the same manner as its CTT and GT counterparts, the investigator should be aware that there exist subtle differences in their meanings and appropriate interpretations (cf. Brennan, 2006).

Recommendations

In sum, reliability may be defined as the consistency of measurement instrument scores across replications of that measurement procedure. A number of statistics for estimating an instrument's (unknown) true reliability are available, including the many reliability coefficients offered by CTT, GT's generalizability coefficients, and the conditional standard errors of measurement yielded by CTT, GT, and IRT. For any particular implementation of a measurement instrument, each of these statistics will be associated with particular strengths and weaknesses associated with (a) the fit of the observed data to the proposed measurement model; (b) practical considerations, such as the sample size required to produce stable estimates (with the IRT statistics requiring somewhat larger samples); and (c) the relevance of the statistic to the applied setting in which it has been used.

With regard to the last consideration, several points are worthy of mention. First, it should be remembered that there exist sometimes-subtle differences between the various reliability and generalizability coefficients developed within the CTT and GT frameworks. In that connection, some coefficients will not be suitable for some intended purposes and populations. A test–retest reliability coefficient, for example, would not be the ideal estimate of the precision of a measure of mood, a construct that will itself vary over time, resulting in changes in observed scores that are unrelated to measurement error. Second, statistics developed within the CTT and GT frameworks are generally sample (e.g., population, D-study design, item) dependent. Third, the standard errors of measurement

yielded by the IRT test information function reflect the restricted configuration of measurement error that is addressed by internal consistency estimates of reliability, and should be interpreted accordingly (AERA, APA, & NCME, 1999). Finally, the investigator should be aware that the particular mathematical function used within the IRT framework to relate item parameters to the observed score can influence the estimated standard errors of measurement (AERA, APA, & NCME).

Does an assessment of the available methods' overall strengths and weaknesses allow for more specific recommendations for practice? The 1999 volume *Standards for Educational and Psychological Testing* (the *Standards*), which was jointly published by the AERA, APA, and NCME, provides some guidance on the most appropriate coefficient for a small set of specific testing purposes (e.g., it recommends that when a measurement instrument is designed to reflect rate of work, a test–retest or alternate-forms coefficient should be used), and its authors emphasize the increasing importance of precision as the potential consequences of measurement error grow in importance (e.g., as in a setting where a single score is used to make decisions about admission to graduate school). In general, however, the *Standards* provides no "cookbook" recommendations regarding the type of reliability evidence that should be sought, nor the level of precision that should be attained. We agree with the authors' assessment that:

> There is no single, preferred approach to quantification of reliability. No single index adequately conveys all of the relevant facts. No one method of investigation is optimal in all situations, nor is the test developer limited to a single approach for any instrument. The choice of estimation techniques and the minimum acceptable level for any index remain a matter of professional judgment. (AERA, APA, & NCME, 1999, p. 31)

Of course, for such judgment to be apt, the investigator must be conversant with the various approaches. Moreover, it is hoped that the investigator will possess sufficient technical resources such that the choice of reliability evidence will be made solely based on the methods' relative strengths and weaknesses and not based on his or her ability (or lack thereof) to enact the different techniques. We hope that our necessarily brief overview of the available methods will spur the reader to pursue any needed additional education on their derivation, computation, and interpretation.

Validity

Brennan (2006) provides an excellent and informative overview of the evolution of measurement theory, and in particular, of historical developments in theoretical models of validity (see also Thompson & Daniel, 1996).

Brennan notes that earlier conceptualizations of validity involved multipartite models that focused on defining specific aspects of validity (e.g., the content, predictive, concurrent, and construct validities defined in the APA's 1954 *Technical Recommendations for Psychological Tests and Diagnostic Techniques*), but emphasizes that more recently theoretical developments have focused on more unified conceptualizations of validity (e.g., Messick, 1988b, 1989) that lend themselves to consideration of multiple means of accumulating evidence relevant to instrument validation.

In an influential 1989 work, Messick defines validity as follows:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment.… [It] is an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use. Hence, what is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators—inferences about score meaning or interpretation and about the implications for action that the interpretation entails. (Messick, 1989, p. 13)

Hence, in ascribing validity to a behavioral measurement process, Messick (1989) focuses on the importance of both (a) establishing the soundness of inferences drawn from the use of the measurement instrument and (b) considering the potential consequences of those inferences. In a separate work, Messick (1988a) offers a four-faceted question that he suggests as a guide for those interested in the process of evaluating validity in the context of behavioral measurement. He asks the potential user of the instrument to consider:

> What balance of evidence supports the interpretation or meaning of the scores; what evidence undergirds not only score meaning, but also the relevance of the scores to the particular applied purpose and the utility of the scores in the applied setting; what rationales make credible the value implications of the score interpretation and any associated implications for action; and what evidence and arguments signify the functional worth of the testing in terms of its intended and unintended consequences. (Messick, 1988a, p. 5)

The latter two facets of Messick's question in particular have engendered discussions of the value of assessing the social (vs. scientific) consequences of measurement (see, e.g., Lees-Haley, 1996; Messick, 1995); they are subjects rife with their own ethical complexities but which fall largely outside the scope of the present chapter. The first two facets of Messick's question, however, are very pertinent to the subject of ethical measurement choices in the research context, and will be addressed below in turn.

First: What balance of evidence supports the interpretation or meaning of the scores?

Score Interpretation

In some applied settings—in recognition of the imperfect nature of behavioral measurement—scores resulting from measurement instruments are not viewed as being perfectly related to the trait or behavior being "measured," but rather, are used to generate hypotheses that are left open to rejection on further investigation. For example, the wise and ethical practicing clinical psychologist would not rely solely on the particular scores resulting from a Rorschach inkblot test to make a definitive diagnosis of psychotic disorder. Instead, the psychologist would consider such scores in the context of a wealth of additional information, such as unstructured interview data, behavioral observations, and records review. When behavioral measurement tools are used in this fashion—as a hypothesis-generating mechanism in the context of an in-depth, individualized assessment paradigm—the potentially negative consequences of imperfect measurement can be dramatically minimized.

In other applied settings, however, scores resulting from behavioral assessment tools are indeed expected to provide a relatively direct index of the trait or behavior being measured, and scores are used in a fashion that is consistent with that expectation. Many tests administered in educational settings, for example, generate scores that are presumed to provide a highly valid index of intellectual and educational aptitude and/or achievement (e.g., end-of-grade testing; the Scholastic Aptitude Test [SAT]). Such scores are used sometimes alone, or in concert with limited additional information, to make consequential decisions about student services, placement, and progression. Readers are encouraged to refer to the chapter written by Cizek and Rosenberg (Chapter 8, this volume) for a discussion of the ethical considerations relevant to these so-called "high-stakes" assessment situations.

In the research setting, and especially in studies that use quantitative methods, assessment tools are used to measure the behavioral constructs under investigation, and the particular scores resulting from these tools are construed as reflecting the level or degree of the trait or behavior being measured. Scores resulting from behavioral measurement tools administered to research participants are generally not of "high stakes" to the participating individuals, in the sense that scores are most often not shared either with the participant or with other decision makers in the participant's life, limiting their potential consequence to the individual. Nevertheless, when investigators apply inferential statistical methods to a full sample of such scores, the particular scores observed will obviously have a critical impact on conclusions drawn about the phenomena

of interest in the population sampled. Hence it is important to establish that the interpretation or assigned meaning of the scores is reflective of the level of the construct being measured.

Kane (2006) provides a highly useful framework for critically evaluating the argument that a particular measurement instrument produces scores that are appropriate for construct-relevant interpretations. Kane asserts that the first step in ensuring interpretable and meaningful scores is to obtain evidence relevant to the scoring of the instrument.

Scoring

Kane (2006) recommends that the user of a behavioral measurement instrument ensure that (a) the rubric used for scoring is appropriate, (b) the rules for scoring are implemented as specified during test construction, and (c) the scoring is unbiased. He notes that many forms of scoring-related evidence could serve to undermine the proposed score interpretations, including, for example, poor interrater agreement, evidence of inadequate training of scorers or raters, and the failure of scoring rules to include relevant criteria. Kane also notes that if a statistical model is used in scaling, it is important to empirically verify that the selected model is a good fit to the observed data.

In making this last point, Kane (2006) focuses primarily on the scaling of scores in the context of standardized testing programs (like the SAT). Under many circumstances, however, new (true or latent) scores are generated by the data analyst when an inferential statistical model is fit to the raw (or scaled) scores resulting from a measurement instrument. The appropriateness and fit of such models are key factors in assessing the validity of the resulting scores. We will return to a discussion of the ramifications of noncontinuous scoring, in particular, on the estimation of measurement models in a later section.

Generalization

The second step Kane (2006) recommends in developing an argument for the interpretability and meaningfulness of scores is provided in the language of GT. In particular, Kane advises that the investigator establish that the observed score to universe score generalization is appropriate for the present use of the instrument. Kane suggests that the investigator first evaluate whether the investigator's sample of observations is representative of the currently defined universe of generalization. Paraphrasing an argument made in an earlier article (Kane, 1996), he opines, "If a serious effort has been made to draw a representative sample from the universe of generalization, and there is no indication that this effort has failed, it would be reasonable to assume that the sample is representative" (p. 35). Concomitantly, Kane suggests that the investigator assess whether the sample size of the present measurement procedure is large

enough to compensate for sampling error. He notes that examination of D-study evidence can point to the presence of problematically large random sampling errors for one or more facets.

Extrapolation

Kane's (2006) extrapolation step involves assurance that the universe score established in the previous step is meaningfully related to the target construct. Such assurance can be obtained using multiple analytic and empirical results, including evaluation of (a) the extent to which the measurement instrument contains items or tasks that are as representative as possible of the construct being assessed (Kane notes that standardization can minimize error, but with the tradeoff that standardized instruments may be associated with a universe of generalization that does not always adequately sample the target domain); (b) *face validity*, or the extent to which the relevance of the measurement instrument to the proposed construct interpretation is apparent to the research participant; (c) *criterion validity*, or the extent to which observed scores on the measurement instrument correlate with scores on a clearly valid criterion measure; and (d) *convergent validity*, or the extent to which observed scores on the measurement instrument correlate with scores on other (perhaps established) measures that seek to tap the same (or a similar) construct.

Implication

The final step in Kane's (2006) framework for appraising whether scores are appropriate for construct-relevant interpretations involves evaluating whether the construct score on the measurement instrument is appropriately linked to the verbal description of that score, and to any implications created by that label. For example, evidence that an achievement-related construct score varies across racial/ethnic groups consisting of members who are otherwise very similar with regard to intellectual ability and educational background would raise serious doubts about the associated measurement procedure's validity.

Threats to Validity

Kane (2006) also urges the investigator to rule out two major threats to the appropriateness of scores to construct-relevant interpretations. The first threat is identified as *trait underrepresentation*, which occurs when a measurement process undersamples the processes and contexts germane to the construct of interest, possibly leading to an overly restrictive universe of generalization. In that connection, Cook and Campbell (1979) and Messick (1989), among others, have emphasized the value of using multiple methods of assessment. The second threat to validity considered by Kane is *irrelevant variance* (vs. random error; also known as *systematic error*), which

is present in the scores derived from a measurement instrument to the extent that those scores are systematically affected by processes that are unrelated to the construct of interest (e.g., rater bias). Multimodal assessment may also minimize irrelevant variance (Messick, 1989).

Score Relevance

What evidence undergirds not only score meaning but also the relevance of the scores to the particular applied purpose and the utility of the scores in the applied setting? The second facet of Messick's (1988a) question may be viewed in part as addressing the question of *generalizability*. In particular, the investigator might ask of the measurement instrument under consideration: Was the instrument originally developed—and validated—for the population and applied purpose that will be the focus of the proposed research? If the answer is no, then the ethical investigator must shoulder the responsibility of seeking out evidence that the measure does operate as intended in the population, and for the purpose, of interest. If such evidence is not available, then he or she should be prepared, as is suggested in the 2002 APA ethics code, to emphasize in the research report the potential limitations of the measurement process and the associated inferences and interpretations.

The procedures outlined in the previous section may be used to validate a measurement instrument in a new population and/or for a novel applied purpose. Those who wish to provide statistical evidence of generalizability may also take advantage of methods for establishing *measurement invariance*. From this perspective, a measurement instrument is considered to be invariant, or generalizable, across populations if participants from different populations who possess the same level of the construct of interest have the same probability of attaining a given score on the instrument (Mellenbergh, 1989). In that connection, latent variable models within both the confirmatory factor analysis and item response theory traditions allow the investigator to evaluate whether fixing parameters relating observed scores to latent variables to be equal across populations results in a significant decrement in model fit (cf. Meade & Lautenschlager, 2004).

Recommendations

In sum, the evaluation of the validity of a measurement process will involve the collection of evidence regarding the measurement instrument's proposed interpretation and use in the population and setting under investigation. Such evidence will likely include information relevant to scoring, generalization, extrapolation, and implication inferences, and will involve assessment of the extent to which the measurement instrument is invariant across populations of interest. It is important to

note that a measurement application that cannot be demonstrated to be adequately reliable will be unlikely to yield sufficient validity evidence (AERA, APA, & NCME, 1999).

Noncontinuous Scores

Many of the methods described above, and especially those drawn from the CTT and GT traditions, are appropriate for the evaluation of measurement instruments that produce continuous scores. However, many forms of behavioral measurement involve classifications or scale types that may fail to produce continuous scores. The investigator who is considering the use of such instruments will encounter at least two decision points in his or her work.

The first question regards whether the noncontinuous nature of the resulting scores calls for different strategies for the evaluation of the instrument's reliability. The answer to this question is likely yes, and that new strategies should especially be considered when scores capture discrete group-membership information. Haertel (2006) provides a useful overview of specialized indices of reliability that are appropriate for either (a) continuous scores that are used to make categorical decisions (as when, e.g., a test is scored and then assigned a value of "pass" or "fail"), or (b) measurement procedures that directly generate classifications into a set of discrete categories. Methods appropriate for classifications involving the comparison of a continuous score with a cut score (or set of cut scores) include Livingston's k^2, Brennan and Kane's Φ and $\Phi(\lambda)$, and Cohen's κ, among many others. Blackman and Koval (1993) provide a discussion of multiple reliability indices that consider the extent of consistency across raters when a measurement procedure involves direct classification into categories.

The second issue confronted by the investigator who contemplates a measurement procedure with noncontinuous outcomes involves consideration of whether inferential measurement models that traditionally use continuous scores may be ethically applied to noncontinuous scores. Most behavioral scientists would acknowledge that few behavioral measurement instruments are possessive of an interval or ratio scale of measurement, and yet application of measurement models and estimation procedures that rely on continuous, interval-level measurement is very common, even when ordinal-level (ordered-categorical) scores have been observed. The appropriateness of such practice has been debated vigorously over the years, both on philosophical and applied grounds (see, e.g., Marcus-Roberts & Roberts, 1987; Michell, 1986; Townsend & Ashby, 1984). For the researcher interested in applying a particular measurement model

to ordinal-level data, both theoretical work and simulation studies, which investigate the behavior of statistical procedures when certain assumptions are violated, can be very informative.

For example, application of the common linear confirmatory factor analysis (CFA) model assumes that observed scores are continuous, or at least interval-level, in measurement. Multiple Monte Carlo simulation studies have addressed the impact within traditional CFA of ordinal-level measurement on the estimation of parameters linking latent variables to observed scores. Wirth and Edwards (2007) note that although results from some studies (e.g., DiStefano, 2002; Dolan, 1994) are suggestive that traditional maximum-likelihood estimation with adjustment might produce acceptable results when the number of ordered categories is five or greater, other findings (e.g., Cai, Maydeu-Olivares, Coffman, & Thissen, 2006) indicate that caution is warranted when applying traditional models to categorical data, even when single-moment adjustments are made.

Fortunately, as Wirth and Edwards (2007) report, multiple statistical frameworks have been developed to accommodate categorical data, among them item factor analysis (IFA). The authors acknowledge the lure of the application of traditional methods to ordinal-level data, especially in light of IFA models' greater complexity and the larger sample sizes typically required to produce stable results. Nonetheless, they recommend that investigators favor IFA when either (a) the number of response categories is fewer than five and/or (b) the present measurement procedure has not yet been well validated in the population of interest. Although they do not uniformly object to the use of traditional methods when measurement procedures produce more than five response categories, in such situations, they strongly encourage the researcher to verify the consistency of results of traditional techniques—obtained using a variety of estimation methods—with those obtained using IFA.

Thus, application of measurement models and estimation procedures that rely on continuous scores to ordinal-level data can potentially produce misleading results. Ideally, investigators will choose measurement models that are appropriate for the level of measurement attained by a particular measurement procedure. If the procedure's level of measurement does not fully satisfy the requirements of a statistical method, it is incumbent on the investigator to critically evaluate the relevant analytical and simulation study findings before applying the method. If the investigator proceeds, she or he should clearly delineate the potential limitations of the method, as applied to data with the observed characteristics, in the research report. Note that although the present discussion has been focused on measurement models, these points are equally relevant to the application of explanatory statistical models such as ANOVA and linear regression to ordinally scaled dependent variables.

We turn next to a review of current standards for the reporting of measurement strategies and associated analyses.

Scientific and Ethical Reporting Standards

Practitioners increasingly are urged—even required—to use evidence-based decision making when choosing interventions and treatments. Such decision making requires access to relevant research that is reported in a manner that allows for an evaluation of its strengths and limitations. To ensure that research reports routinely include all the information necessary for weighing the evidence produced, a number of professional organizations have articulated reporting standards. These standards cover all facets of research—from conceptualization to design and analysis—and to varying degrees they address the integrity of the measures and measurement strategies used.

Origins in Biomedical Research

Perhaps the most organized efforts at standardizing research reports and ensuring that they include all the information bearing on the strength and limitations of a study have been within the context of biomedical research. At least two statements and accompanying checklists are primarily aimed at guiding the reporting of biomedical research findings: the Consolidated Standards of Reporting Trials (CONSORT) statement, which applies to randomized trials, and the Transparent Reporting of Evaluations with Nonrandomized Designs (TREND) statement, which applies to nonrandomized evaluation studies. The standards prescribed by these statements largely focus on accounting for research participants and on description of the intervention or treatment, procedure for assigning participants to condition, and methods involved in statistical inference; however, both touch on measurement. Item 6a in the CONSORT statement, for example, prescribes "clearly defined primary and secondary outcome measures" (Altman et al., 2001, p. 669). The explanation associated with this item provides several guidelines for reporting. First, primary outcomes should be clearly identified and distinguished from secondary outcomes. Second, when scales or instruments are used, "authors should indicate [their] provenance and properties" (p. 669). Finally, the statement urges the use of "previously developed and validated scales" (p. 669; see also Marshall

et al., 2000). Item 6b in the CONSORT statement refers to steps taken to improve the reliability of measurement and indicates that the use of multiple measurements or assessor training should be described fully in the report. Although the CONSORT guidelines have been endorsed by more than 150 journals, general adherence by authors and enforcement by journal editors have not been uniform (Barbour, Moher, Sox, & Kahn, 2005).

The TREND statement was patterned after the CONSORT statement but adds items relevant to research in which participants are not randomized to condition (Des Jarlais, Lyles, & Crepaz, 2004). To the CONSORT measurement recommendations, the TREND statement adds only the prescription that "methods used to collect data" (e.g., self-report, interview, computer-assisted) should be described.

Before turning to standards for reporting behavioral research, we note one additional set of standards that is focused almost entirely on measurement: the Standards for the Reporting of Diagnostic Accuracy Studies (STARD; Bossuyt et al., 2003). The STARD checklist includes 25 items, of which approximately half concern the reporting of measurement methods and the analysis of the effectiveness of the "index test" to be used for diagnoses. Prescriptions to report information about the reference standard, reproducibility, and accuracy of classification—for the sample as a whole and for subgroups of interest—reflect the importance of a clearly articulated evidence base for tests that will be used by practitioners. Use of the STARD checklist is encouraged by more than 200 biomedical journals.

Standards for Behavioral Science

In behavioral science, only recently have formal statements of reporting standards been published, and to date, these have not been formally endorsed by specific journals. As a result, the primary audience for such standards is manuscript authors. Adherence to the standards is voluntary and uneven; hence they currently appear to function more as recommendations than standards as strictly defined.

In 1999, the APA Task Force on Statistical Inference published a set of guidelines for the reporting of statistical methods (Wilkinson & the Task Force on Statistical Inference, 1999). Building on an earlier report by the International Committee of Medical Journal Editors (Bailar & Mosteller, 1988), the APA Task Force offered guidelines for reporting the investigator's selected methods and associated results and for drawing appropriate conclusions. Unlike the earlier report, the report of the Task Force included a lengthy section on measurement in which the authors offer guidance on describing measurement procedures. With regard to variables, the Task

Force asserts, "Naming a variable is almost as important as measuring it. We do well to select a name that reflects how a variable is measured" (p. 596). With regard to the use of a questionnaire measure, the Task Force urges authors to "summarize the psychometric properties of its scores with specific regard to the way the instrument is used in the population" (p. 596). *Psychometric properties* were defined as "measures of validity, reliability, and any other qualities affecting conclusions" (p. 596). Finally, the Task Force proposes that authors provide detail about how the measures were used, recommending that authors "clearly describe the conditions under which measurements are taken (e.g., format, time, place, personnel who collected the data)" (p. 596). Of particular concern to the Task Force were aspects of the measurement procedure that might introduce bias, and they instruct authors to describe measures taken to reduce or eliminate potential biases.

A similar report was produced by the American Education Research Association's (AERA, 2006) Task Force on Reporting of Research Methods in AERA Publications. To the information requested by the APA Task Force, the AERA Task Force adds with reference to the description of measures used in research that "information on access to these surveys, instruments, protocols, inventories, and guides should be specified" (p. 36). In addition, prescriptions for detailing steps taken to develop new measures or to classify research participants using scores are offered in a "Measurement and Classification" section.

The move from guidelines and suggestions to potential standards for reporting of behavioral science is more apparent with the efforts of the APA Publications and Communications Board Working Group on Journal Article Reporting Standards (the Working Group, 2008). The Working Group began by consolidating the CONSORT, TREND, and AERA standards described earlier and then added new reporting recommendations. Like the CONSORT and TREND standards, the resultant recommendations refer to all aspects of a report of empirical research. In a section labeled "Measures and Covariates," the Working Group recommends that reports of new research include (a) definitions of all variables, including primary and secondary variables and covariates (to include mention of variables on which data were gathered but not analyzed for the report); (b) measurement methods, including reference to any training of individuals who administered measures and consistency between measures when administered more than once; and (c) "information on validated or ad hoc instruments created for individual studies, for example, psychometric and biometric properties" (p. 842). Unfortunately, perhaps because of the intended generality of the proposed standards, neither details nor examples are provided for the Working Group's recommendations.

The APA Working Group was one of seven working groups that contributed to the production of the sixth edition of the *Publication Manual of the*

American Psychological Association (APA, 2010), and its recommendations for research reports are reflected in that influential document. In fact, the content of the "Measures and Covariates" item from the Working Group report was carried forward into the *Manual* without elaboration. Notably, in a chapter on manuscript structure and content, the *Manual*'s recommendations are referred to as reporting "standards," although authors are encouraged to "balance the rules of the *Publication Manual* with good judgment" (p. 5).

To date, perhaps the most concrete, prescriptive, and exhaustive set of recommendations addressing the development, evaluation, and appropriate documentation of behavioral measurement instruments is provided by the 1999 *Standards* (AERA, APA, & NCME). Although the primary aim of the *Standards* was to provide guidance to those involved with educational, personnel, and program evaluation testing applications (see Cizek & Rosenberg, Chapter 8, this volume), the recommendations are sufficiently general that they are relevant for the broader applied measurement and behavioral science research community. In summarizing the overall purpose and intended audience of the *Standards*, the authors advocate that "within feasible limits, the relevant technical information be made available so that those involved in policy debate may be fully informed" (p. 2).

With regard to reporting associated with the precision of a measurement procedure, the *Standards* emphasizes that "general statements to the effect that a test is 'reliable' or that it is 'sufficiently reliable to permit interpretations of individual scores' are rarely, if ever, acceptable" (AERA, APA, & NCME, 1999, p. 31). For the selection, evaluation, and reporting of data germane to the assessment of a measurement instrument's reliability, the authors provide the following guidance:

- For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported (Standard 2.1, p. 31).
- The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original sale units and in units of each derived score recommended for use in test interpretation (Standard 2.2, p. 31).
- When test interpretation emphasizes differences between two observed scores of an individual or two averages of a group, reliability data, including standard errors, should be provided for such differences (Standard 2.3, p. 32).
- Each method of quantifying the precision or consistency of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to

select examinees for reliability analyses and descriptive statistics on these samples should be reported (Standard 2.4, p. 32).

- A reliability coefficient or standard error of measurement based on one approach should not be interpreted as interchangeable with another derived by a different technique unless their implicit definitions of measurement error are equivalent (Standard 2.5, p. 32).
- Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score (Standard 2.14, p. 35).

Additional reliability standards address reporting for particular testing applications (e.g., tests designed to reflect rate of work, tests scored by raters, tests with both long and short forms). The reader is encouraged to consult the *Standards* for annotations and additional details.

The *Standards* (AERA, APA, & NCME, 1999) additionally addresses the reporting of validity evidence. Recommendations particularly relevant to the use of an existing measure in a research context include:

- If a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary (Standard 1.4, p. 18).
- The composition of any sample of examinees from which validity evidence is obtained should be described in as much detail as is practical, including major relevant sociodemographic and developmental characteristics (Standard 1.5, p. 18).
- When interpretation of performance on specific items, or small subsets of items, is suggested, the rationale and relevant evidence in support of such interpretation should be provided (from Standard 1.10, p. 19).
- If the rationale for a test use or interpretation depends on premises about the relationships among parts of the test, evidence concerning the internal structure of the test should be provided (Standard 1.11, p. 20).
- When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given (Standard 1.12, p. 20).
- When validity evidence includes empirical analyses of test responses together with data on other variables, the rationale

for selecting the additional variables should be provided. Where appropriate and feasible, evidence concerning the constructs represented by other variables, as well as their technical properties, should be presented or cited. Attention should be drawn to any likely sources of dependence (or lack of independence) among variables other than dependencies among the construct(s) they represent (Standard 1.14, p. 20).

- When validation relies on evidence that test scores are related to one or more criterion variables, information about the suitability and technical quality of the criteria should be reported (Standard 1.16, p. 21).
- If test scores are used in conjunction with other quantifiable variables to predict some outcome or criterion, regression (or equivalent) analyses should include those additional relevant variables along with the test scores (Standard 1.17, p. 21).
- When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported (Standard 1.18, pp. 21–22).

Other standards not presented here address the reporting of validity evidence for applications such as treatment assignment, use of meta-analytic evidence for instrument validation, and consequences of testing. Again, the reader is encouraged to consult the 1999 volume for a comprehensive treatment of reporting recommendations.

Conclusion

In sum, the 2002 APA ethics code urges the behavioral scientist to select instruments that are useful and properly applied and whose reliability and validity have been established in the population of interest. The code provides no recommendations for procedures for establishing reliability and validity; in this chapter, we have attempted to provide more specific guidance. The code further prescribes that the scientist report evidence supportive of his or her instrument selection. Multiple professional organizations have articulated standards that address the appropriate reporting of measurement strategies; earlier efforts originated in the field of biomedical research and are in general neither detailed nor comprehensive enough to address the varied measurement concerns associated with

behavioral sciences research. In recent years, formal statements of reporting standards for research have been published for the behavioral sciences in particular, but these statements, including the APA's 2010 *Publication Manual*, reflect barely expanded measurement sections. To date, perhaps the most usefully prescriptive and exhaustive set of reporting recommendations for the behavioral sciences is provided by the 1999 *Standards* (AERA, APA, & NCME). It is unfortunate that the APA's *Publication Manual*, which is so widely used and readily available to those in the field, does not contain a more comprehensive measurement section. Indeed, awareness of ethical measurement practice in the behavioral sciences might be heightened if that section were expanded in future editions. In our view, the accountability associated with the reporting of evidence supportive of an instrument's implementation can only serve to improve adherence to ethical conduct.

We believe that a fair share of the burden involved in assuring the ethical use of behavioral measurement instruments falls to the instrument developer. Ideally, the developer will provide evidence of the reliability and validity of the instrument in the clearly defined population and applied setting for which it was designed, and—anticipating future uses to the extent possible—both provide reliability data for alternative populations and warn potential users against unsupported interpretations (AERA, APA, & NCME, 1999). Unquestionably, however, the research user of a behavioral measurement instrument bears a significant responsibility for ensuring that his or her choice of instrument has led to valid inferences.

The investigator should be aware that the inferences derived from measurement choices occur at all stages of the research process. Accordingly, the investigator following ethical practice will evaluate validity evidence relating to scoring, generalization, extrapolation, and implication inferences (Kane, 2006) and the generalizability of the instrument to the present population. The researcher will also be conversant with the various statistical frameworks for estimating instrument precision and will evaluate the particular forms of reliability evidence that are most relevant to her or his proposed use of the instrument, taking care to consider the fit of any measurement models used to the observed data. In analyzing study data, the investigator will choose measurement models that are appropriate for the level of measurement attained by a particular measurement procedure, taking into account model assumptions and relevant analytical and simulation study findings; furthermore, if latent variable models are used, the names given to latent constructs will be selected and explained with great care. Finally, the investigator should adhere to the scientific and ethical reporting standards summarized above and refer especially to the *Standards* (AERA, APA, & NCME, 1999) for concrete recommendations relevant to particular measurement strategies. The researcher should discuss in the research report the potential limitations of the measurement

process and its associated inferences and interpretations. At all stages of the research and reporting process, informed professional judgment will be required.

Case Example

Many of the issues we have raised are highlighted in the literature on the design and interpretation of the Implicit Association Test (IAT). We have stressed the importance of assessing the reliability and validity of a measurement instrument in a particular population and applied setting; in the present section, we do not contravene our earlier advice by offering discussion about the properties of the IAT per se, but we do attempt to pinpoint issues that have likely pertained to most, if not all, of the instrument's applied uses.

The IAT is billed primarily as a measure of implicit bias toward a specific group (e.g., an ethnic minority, the elderly). Here, an *implicit* bias refers to one that is beneath awareness and presumably outside conscious control (Greenwald & Banaji, 1995); it may be contrasted with an *explicit* bias, of which the individual presumably is aware and would be able to control if motivated to do so. An intriguing feature of the IAT is that those completing the measure reportedly feel unable to control their implicit biases, even when they realize their responses may be revealing them.

The initial description of the IAT was published in the June 1998 issue of the *Journal of Personality and Social Psychology* (Greenwald, McGhee, & Schwartz). In its most basic form, the IAT is administered by seating the respondent at a computer. Displayed on the monitor are words and/or images, to which the respondent reacts by pressing a key. Most frequently, the reaction involves classifying the word or image into one of two contrasting categories (e.g., *young* vs. *old*, *good* vs. *bad*). Response data reflect the computed latency between the time the word or image appears on the screen and the time at which the respondent presses the key to indicate the correct categorization. The response latencies for different types of categorization are combined to produce an overall score that reflects any bias favoring one group over the other. Imagine, for example, that the IAT was being used to assess bias against the elderly: If the respondent were faster to categorize young faces and positive using the same key and old faces and negative using the same key than they were to categorize young and negative and old and positive, then the respondent would be assumed to possess an unconscious bias in favor of young people and against older people. On the other hand, if the opposite pairing were faster, then the respondent's score would be assumed to reflect

an unconscious bias in favor of older people and against younger people. After its development, the IAT was quickly and widely endorsed by the larger research community, as evidenced by its use in at least 122 research studies published through January 2007 (summarized in the meta-analytic report of Greenwald, Poehlman, Uhlmann, & Banaji, 2009).

The IAT has also captured interest outside academe. The IAT was published on a freely accessible website in October 1998; through Project Implicit, funded by the National Institute of Mental Health and the National Science Foundation, the IAT remains available for self-administration at https://implicit.harvard.edu/implicit, where it is completed approximately 15,000 times each week (with a total of approximately 4.5 million completions since it first appeared online; Project Implicit, n.d.). The measure is also regularly featured in the popular media (e.g., Chedd, 2007; Thompson, 2009; Tierney, 2008; Vedantam, 2005), where it is described using statements such as, "The tests get to the bottom of our true inclinations" (Thompson, para. 3).

The rare recognition and acceptance of the IAT by the larger public has led to scrutiny of the measure that is somewhat uncommon for instruments developed primarily for research purposes. The result of this scrutiny is a growing literature questioning the reliability and validity of implementations of the IAT. These questions focus primarily on three broad concerns.

The first concern is that the reliability evidence associated with the IAT may not be sufficiently consistent and strong to warrant the relatively unqualified acceptance the measure has enjoyed. To date, the small amount of information offered on the precision of the IAT has been in the form of test–retest reliability coefficients. Because the IAT is purported to tap an individual difference, its associated short-term test–retest coefficients should theoretically be high, perhaps in the neighborhood of .80 or greater. However, the range for 1-week to 1-month coefficients has typically ranged from .50 to .70 (e.g., Bosson, Swann, & Pennebaker, 2000). Such reliability estimates reflect reasonably high levels of measurement error and are furthermore not consistent with the idea that the IAT taps a stable characteristic.

The second concern is that the observed scores produced by the IAT may not be sufficiently valid measures of the construct of interest. The validity of the IAT as a measure of unconscious bias has been questioned since the measure was first introduced, and indeed, its developers offer access to more than 50 articles addressing validity concerns at http://faculty.washington.edu/agg/iat_validity.htm. Multiple forms of evidence appear to cast doubt on the appropriateness of construct-relevant interpretations. For example, evidence regarding the strength of association between implicit and explicit measures of the same bias is inconsistent, yet theoretical explanations for observed associations appear to adapt, to some

degree, to the observed data: Although correlations between measures of extrinsic bias and IAT measures of intrinsic bias vary widely across studies, a correlation produced in any single research study—whether high or low—tends to be interpreted in a manner that is favorable to application of the IAT. Indeed, in the seminal paper on the IAT (Greenwald, McGhee, & Schwartz, 1998), the authors report that two explicit measures of bias correlated with each other at an approximate r of .60, whereas the same measures correlated with the IAT measure of implicit bias at $r = .25$. Although these findings might reasonably be interpreted as a failure to demonstrate convergent validity, the authors argue that they are instead an important demonstration of discriminant validity; however, the authors' provided evidence of criterion validity comes in the form of prediction by IAT scores controlling for explicit measures. Moreover, as described above, one major threat to the validity of a behavioral measurement instrument is irrelevant variance. In that connection, it is unfortunate that a detailed analysis of the IAT aimed at specifying an appropriate measurement model has revealed that variability in IAT scores can be attributed to a number of influences, including a cluster of variables that influences general processing speed (e.g., attention span, hand–eye coordination, mood; Blanton, Jaccard, Gonzales, & Christie, 2006). Questions have also been raised about the validity of the difference scores computed as part of the IAT. Although these issues have been addressed in revised forms of the test (e.g., Blanton et al.; Greenwald, Nosek, & Banaji, 2003; Olson & Fazio, 2004), they are relevant for a significant portion of the existing literature on IAT-assessed implicit cognition. Although a recently published meta-analysis suggests that IAT scores offer incremental validity over scores on traditional self-report measures (Greenwald et al., 2009), questions about the validity of IAT applications persist (e.g., Arkes & Tetlock, 2004).

A third set of concerns stems from the consequences of the ready public availability of the IAT. Upwards of 2,000 people per day complete an IAT at the Project Implicit website; multiple forms of the test are available on the website, including versions that purportedly reveal automatic preferences for "light-skinned" versus "dark-skinned" faces and for disabled versus abled individuals. At the conclusion of each test, the respondent receives a brief feedback statement (described as a "score;" e.g., "Your data suggest a moderate automatic preference for Young people compared to Old people") and is provided the opportunity to view a frequency table that provides the percentage of Internet respondents that received each possible "score."

As we have shown, however, the reliability and validity of IAT scores—although sufficient to support continuing research and development—are not strong. The Project Implicit team has been careful to avoid overstating the validity of the measure, stating in the website's FAQ that "these tests are not perfectly accurate by any definition of accuracy." But media reports and other websites that steer people to the Project Implicit website are not as

careful, and it seems unlikely that the average test-taker would consult the large amount of information about implicit cognition and the IAT offered by the Project Implicit team in a separate section of their website. Hence it is potentially misleading, and perhaps even harmful, for the website to communicate to the respondent that his or her IAT performance reflects the presence or absence of unconscious bias. A serious concern is the respondent's interpretation of his or her website-provided feedback and the potential consequences of that interpretation. For test-takers who prefer to view themselves as unprejudiced, an unquestioned IAT "score" that indicates otherwise could cause distress; for any test-taker, shared feedback could have social repercussions. Finally, although the IAT seeks to tap a presumably stable (and unconscious) individual difference, it is unclear how an individual's use of the Project Implicit website might impact his or her later responses to the IAT in the context of participation in a research study.

In sum, the IAT is an intriguing instrument that has captured the attention of many in the behavioral sciences research community and the interest of the media and members of the general public alike. Unfortunately, the widespread use of the IAT raises important ethical questions. Although ethical practice would suggest the use of instruments whose reliability and validity have been supported in the population tested, such evidence is not currently sufficient even for controlled research applications of the IAT, and to our knowledge, it is virtually nonexistent for the Internet-based population of IAT test-takers. Fundamental questions especially persist regarding the appropriateness of the IAT for construct-relevant interpretations. The impact of such questions on the validity of experimental findings is clear. Of equal concern are questions regarding the usefulness and potential impact, at both the individual and societal levels, of self-interpretation of IAT performances on the publicly available forms of the test.

As is illustrated here, basic concerns about how behavioral measures are described, used, and administered are more the norm than the exception. The IAT is a particularly useful case example because it illustrates these concerns as they play out in both the research context and in popular culture. Although the translation of somewhat arcane processes such as implicit cognition into a form that resonates with the public is a commendable goal, the IAT does serve as an example of the challenges involved.

References

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., … Lang, T. (2001). The Revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine, 134,* 663–694.

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35,* 33–40.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: Author.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060–1073.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839–851.

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test?" *Psychological Inquiry, 15,* 257–278.

Bailar, J. C., III, & Mosteller, F. (1988). Guidelines for statistical reporting in articles for medical journals: Amplifications and explanations. *Annals of Internal Medicine, 108,* 266–273.

Barbour, V., Moher, D., Sox, H., & Kahn, M. (2005). Standards of reporting biomedical research: What's new? *Science Editor, 28,* 4.

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods, 5,* 370–379.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Blackman, N. J.-M., & Koval, J. J. (1993). Estimating rater agreement in 2 x 2 tables: Corrections for chance and intraclass correlation. *Applied Psychological Measurement, 17,* 211–223.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the Implicit Association Test: Implications for criterion prediction. *Journal of Experimental Social Psychology, 42,* 192–212.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology, 79,* 631–643.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., … de Vet, H. C. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *Annals of Internal Medicine, 138,* 40–44.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). Westport, CT: Praeger Publishers.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^P tables. *British Journal of Mathematical and Statistical Psychology, 59,* 173–194.

Chedd, G. (Writer and Director). (2007). The hidden prejudice [Television series episode]. In J. Angier & G. Chedd (Executive Producers), *Scientific American Frontiers*. Public Broadcasting Corporation.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Des Jarlais, D. C., Lyles, C., & Crepaz, N. (2004). Improving the reporting quality of nonrandomized evaluations of behavior and public health interventions: The TREND statement. A*merican Journal of Public Health, 94,* 361–366.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9,* 327–346.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5, and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47,* 309–326.

Dunivant, N. (1981). *The effects of measurement error on statistical models for analyzing change: Final report* (Grant NIE-G-78-0071). Washington, DC: National Institute of Education (Educational Resources Information Center Document Reproduction Service No. ED223680). Retrieved from Educational Resources Information Center database.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: American Council on Education and Macmillan.

Feldt, L. S., & Charter, R. A. (2003). Estimating the reliability of a test split into two parts of equal or unequal length. *Psychological Methods, 8,* 102–109.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4–27.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74,* 1464–1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85,* 197–216.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97,* 17–41.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: Praeger Publishers.

Hägglund, G. (1982). Factor analysis by instrumental variables. *Psychometrika, 47,* 209–222.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12,* 38–47.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hoyle, R. H., & Kenny, D. A. (1999). Sample size, reliability, and tests of statistical mediation. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 195–222). Thousand Oaks, CA: Sage Publications.

Huitema, B. E. (1980). *The analysis of covariance and alternatives.* New York: Wiley.

Jöreskög, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57,* 239–251.

Kane, M. (1996). The precision of measurements. *Applied Measurement in Education, 9,* 355–379.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger Publishers.

Koocher, G. P., & Keith-Spiegel, P. (2008). *Ethics in psychology and the mental health professions: Standards and cases* (3rd ed.). New York: Oxford University Press.

Lees-Haley, P. R. (1996). Alice in validityland, or the dangerous consequences of consequential validity. *American Psychologist, 51,* 981–983.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517–548.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28,* 989–1020.

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

Marcus-Roberts, H. M., & Roberts, F. S. (1987). Meaningless statistics. *Journal of Educational Statistics, 12,* 383–394.

Marshall, M., Lockwood, A., Bradley, C., Adams, C., Joy, C., & Fenton, M. (2000). Unpublished rating scales: A major source of bias in randomized controlled trials of treatments for schizophrenia. *British Journal of Psychiatry, 176,* 249–252.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies in establishing measurement equivalence/invariance. *Organizational Research Methods, 7,* 361–388.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127–143.

Messick, S. (1988a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5–11.

Messick, S. (1988b). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50,* 741–749.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin, 100,* 398–407.

Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extra-personal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology, 86,* 653–667.

Project Implicit. (n.d.). Retrieved from http://projectimplicit.net/generalinfo.php

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356–442). Washington, DC: American Council on Education.

Thompson, B., & Daniel, L. G. (1996). Seminal readings on reliability and validity: A "hit parade" bibliography. *Educational and Psychological Measurement, 56,* 741–745.

Thompson, J. (2009). *Project Implicit: Am I racist?* Retrieved from http://www.myfox-chicago.com/dpp/news/project-implicit-am-i-a-racist-dpgo-20091029-jst1256860352012

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560–620). Washington, DC: American Council on Education.

Tierney, J. (2008, November 17). In bias test, shades of gray. *The New York Times*. Retrieved from http://www.nytimes.com/2008/11/18/science/18tier.html

Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin, 96,* 394–401.

Vedantam, S. (2005, January 23). See no bias. *The Washington Post*, p. W12.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12,* 58–79.

Wright, T. A., & Wright, V. P. (2002). Organizational researcher values, ethical responsibility, and the committed-to-participant research perspective. *Journal of Management Inquiry, 11,* 173–185.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: Praeger Publishers.