Psychological Assessment
2003, Vol. 15, No. 4, 446–455

1040-3590/03/$12.00 DOI: 10.1037/1040-3590.15.4.446

The Incremental Validity of Psychological Testing and Assessment: Conceptual, Methodological, and Statistical Issues

John Hunsley
University of Ottawa

Gregory J. Meyer
University of Toledo

There has been insufficient effort in most areas of applied psychology to evaluate incremental validity. To further this kind of validity research, the authors examined applicable research designs, including those to assess the incremental validity of test instruments, of test-informed clinical inferences, and of newly developed measures. The authors also considered key statistical and measurement issues that can influence incremental validity findings, including the entry order of predictor variables, how to interpret the size of a validity increment, and possible artifactual effects in the criteria selected for incremental validity research. The authors concluded by suggesting steps for building a cumulative research base concerning incremental validity and by describing challenges associated with applying nomothetic research findings to individual clinical cases.

The concept of incremental validity is essentially a simple and straightforward one: does a measure add to the prediction of a criterion above what can be predicted by other sources of data? Studying and applying this simple concept in the realm of applied psychology inevitably leads to increased complexity, as an improvement in prediction can be demonstrated in multiple ways, including increased power, sensitivity, specificity, and predictive efficacy of decision-making judgments beyond what is generated on the basis of other data (Haynes & O'Brien, 2000). In clinical contexts, assessment can be conducted for numerous reasons, including diagnosing a disorder or problem, developing a case conceptualization, treatment planning, treatment monitoring, and/or treatment outcome evaluation. Thus, a measure may have incremental validity in some assessment applications but not others. Finally, assuming a measure has been demonstrated to have incremental validity in a specific applied decision-making task, it then becomes important to consider (a) the range of circumstances in which the measure makes an incremental contribution and (b) the cost issues, as the financial and human resource costs associated with the measure must be balanced against the applied value of the validity increment (Yates & Taub, 2003).

Over the past several decades, numerous commentators on testing and assessment practices in a range of applied psychology domains have called for greater attention to the role that incremental validity data should play in determining psychologists' use of interviews, observations, and psychological tests (e.g., Elliott, O'Donohue, & Nickerson, 1993; Mash & Hunsley, 2004; Meyer & Archer, 2001; Stout & Cook, 1999; Widiger & Schilling, 1980; Wiggins, 1973). Despite this, there has been little systematic effort in most areas of applied psychology to evaluate the incremental validity of measures and assessment procedures. Our goal in this article is to present and critically examine key issues that must be considered if these calls for evidence of incremental validity are to be met in a scientifically rigorous manner. We begin our discussion of these issues by providing an overview of the concept of incremental validity in the research and clinical use of psychological tests and assessment procedures (including standardized tests, structured interviews, and observational procedures). We then discuss a number of statistical and measurement issues that influence the validity and interpretation of incremental validity research. We conclude by highlighting the implications of these issues for future research and for furthering research-informed psychological assessment practices.

Incremental Validity: An Overview

During the 1950s, theoretical and applied work on test validity burgeoned. As part of these developments, some psychometricians began to suggest that newly developed tests intended to be used for personnel decisions should demonstrate an ability to add to the prediction of outcomes beyond that which was possible with the best available assessment strategies (Cronbach & Gleser, 1957). Building on this work, as well as on the contributions of Campbell (Campbell, 1960; Campbell & Fiske, 1959) and the recommendations of the American Psychological Association's (1954) Committee on Psychological Tests, it was Sechrest (1963) who first proposed and articulated the concept of incremental validity. He argued that, in addition to evidence for convergent and discriminant validity, a psychological test that was intended for applied use (i.e., academic, clinical, or personnel applications) must yield an improvement in prediction compared with the result derived from using data that are easily and routinely obtained as part of the process of assessment. This requirement presents a rather stringent test of validity, as it requires not only that the prediction of an

Editor's Note. Yosef S. Ben-Porath served as the action editor for this article.—SNH

John Hunsley, School of Psychology, University of Ottawa, Ottawa, Ontario, Canada; Gregory J. Meyer, Department of Psychology, University of Toledo.

Correspondence concerning this article should be addressed to John Hunsley, School of Psychology, University of Ottawa, Ottawa, Ontario, Canada K1N 6N5. E-mail: hunch@uottawa.ca

outcome with a test be better than that obtained by chance but also that the test demonstrate its value in comparison with other relevant sources of information. Minimally, for a test to have true utility in an applied context, Sechrest suggested that the test should demonstrate incremental validity over brief case history information, simple biographical data, and brief interviews. Setting the standards even higher, he further suggested that a test should make a contribution to the predicted outcome over that possible with simpler and less expensive psychological tests. In an earlier discussion of similar issues, Meehl (1959) recommended an additional factor be considered in evaluating the incremental value of a test, namely, the extent to which the increment in prediction is associated with the provision of services that are beneficial to a person being assessed (e.g., does the increment lead to more effective treatment than would otherwise be provided).

The next major reference to the concept of incremental validity appeared in Wiggins' (1973) text *Personality and Prediction: Principles of Personality Assessment.* Adding to Sechrest's (1963) presentation of statistical issues in demonstrating incremental validity, Wiggins explicitly contrasted the value of a personality test when making personnel decisions against base-rate information (e.g., the general frequency of success or turnover in a setting) and provided an equation for calculating the extent to which personnel selection based on test data might improve on random selection and base-rate data. Wiggins cautioned that conclusions about the incremental validity of a test are context specific, as the results obtained with a given base rate may not generalize to a situation in which the base rate is substantially different. Moreover, he explicitly raised the possibility that the incremental validity of a test over other readily available information may be so small that it may not be worth the financial cost associated with the use of the test.

In later editions of her classic text *Psychological Testing,* Anastasi (1988) summarized key issues in incremental validity, succinctly indicating that incremental validity depends on base rates and selection ratio (i.e., the number of candidates to be selected in comparison with the number of applicants) considerations. She concretely demonstrated the effect of selection on validity coefficients for specific base-rate levels and, like Wiggins (1973), urged caution in attempting to generalize across samples with divergent base rates. In particular, she emphasized that situations involving very low base rates (i.e., very rare or very common events) are especially problematic: any appropriate and valid test may be able to demonstrate incremental validity, but the increment is likely to be extremely small. Given that the diagnosis of clinical conditions is likely to occur in the context of disorders with low base rates, she urged that close attention be paid to the financial costs associated with test administration and to the financial and psychological costs accruing from the inevitable false positives that would occur in the clinical context. Consistent with previous presentations of incremental validity in assessment, Anastasi focused on clinical decisions or predictions in the context of nomothetic or group-level situations. She did not discuss the extent to which these nomothetically based methods or criteria for establishing incremental validity were also relevant to the context of idiographic clinical assessment in which decisions or predictions are focused on specific individuals.

Despite the discussion of incremental validity in assessment texts, there was no systematic evaluation of the incremental validity of clinical assessment research until Garb's (1984) influential review of the incremental validity of interview data, biographical data, personality tests, and neuropsychological tests. In general, Garb found that biographical data, Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1943) results, and neuropsychological test results all had incremental validity for assessing the psychological functioning of adults in specific prediction contexts. However, findings for other types of assessment data were not as positive. For instance, the addition of interview videotapes to transcripts did not lead to a validity increment, whereas the Rorschach, Thematic Apperception Test (Murray, 1943), and sentence completion measures produced inconsistent results. Garb also indicated that the widespread acknowledgment of the importance of incremental validity data had not been translated into much actual research on incremental validity. Indeed, in his review, only 32 studies met his inclusion criteria and many commonly used measures and assessment strategies had never been evaluated within an incremental validity paradigm. Finally, Garb noted that the existing research was not cumulative and that little effort appeared to have been made to follow up on promising incremental validity results for a test or assessment strategy.

Since Garb's (1984) review, there have been scores of incremental validity studies in the various domains of psychology, with many different measures used and a variety of concepts and prediction tasks studied. Researchers have examined incremental validity in the context of adding a measure of anxiety sensitivity to personality dimensions in predicting elements of panic attacks (Lilienfeld, 1997), adding self-reports of cognitive ability to the results of neuropsychological tests in predicting disease-related cognitive deficits (Schwartz, Kozora, & Zeng, 1996), and using response latencies to detect dissimulation (Holden & Hibbs, 1995) and predict flight training performance (Siem, 1996). Incremental validity studies have also been designed to address such questions as determining the extent to which test administration time can be optimized by the use of subtests rather than a full test battery (e.g., Woodard & Axelrod, 1995) and examining the best combination of assessment measures to use in a clinical evaluation context (e.g., Lofland, Cassisi, Levin, Palumbo, & Blonsky, 2000). Within this large literature, a number of incremental validity studies have focused specifically on the use of intelligence test data in predicting academic achievement (e.g., Glutting, Youngstrom, Ward, & Ward, 1997; Kline, Snyder, Guilmette, & Castellanos, 1992; Watkins & Glutting, 2000).

To date, of all areas of applied psychology, incremental validity research seems to have had the most impact on the area of personnel psychology. Indeed the research in this area is sufficiently large and focused that meta-analytic studies have been conducted to summarize the results of decades of research on issues such as employment interview evaluations (Huffcutt, Roth, & McDaniel, 1996), the use of informant ratings to assess job performance (Conway, Lombardo, & Sanders, 2001), and personnel selection strategies (Schmidt & Hunter, 1998). Attention to, and reliance on, incremental validity research appears to be firmly ensconced in the personnel psychology area; unfortunately, the progress made in addressing issues of incremental validity in this area has not been paralleled by the same degree of progress in other areas of applied psychology (cf. Johnston & Murray, 2003). In many respects, this differential progress is not surprising because there are a number of clearly defined applied prediction tasks in personnel research (e.g., job success, job turnover) and strong financial contingencies for maximizing the prediction of those outcomes (e.g., corporate expenses for applicant screening).

Incremental Validity: Conceptualization and Research Design Considerations

As originally presented by Sechrest (1963) and Wiggins (1973), incremental validity was conceptualized as an applied form of validity, inasmuch as the purpose of incremental validity was to provide evidence pertinent to improving on decision making and prediction tasks. Within this general frame, there are three overlapping but relatively distinct conceptualizations of incremental validity research evident in the psychological literature, including the incremental validity of testing instruments, of test-informed clinical inferences, and of new measures. However, before describing and illustrating these three approaches to incremental validity, we note that the concept of incremental validity has also been presented as a generic form of validity that describes the ability of a measure (or a professional) to predict a variable of interest beyond what is possible with other data (e.g., Dawes, 1999; Haynes & Lench, 2003). Some researchers have examined the incremental validity of a measure for conceptual, rather than purely applied reasons, such as providing evidence of construct validity for a measure (Boland & Cappeliez, 1997; Hunsley, 1987; Judge, Erez, Bono, & Thoresen, 2002; Meyer, 2000), demonstrating the relevance of a construct to a specific assessment task (Donnay & Borgen, 1999), or model testing (James & Hunsley, 1995). As these approaches to incremental validity are not as directly relevant to the question of the utility of assessment, we do not focus on them in this article.

Incremental Validity of Testing

In this approach to examining incremental validity, nomothetic analyses are conducted in which information from a new source of data (such as a test scale or an observational coding system) is examined in terms of its contribution to improving on the prediction of a clinically relevant criterion (e.g., diagnosis, adjustment, treatment attendance, or treatment outcome). The focus in this type of research is on the value of adding new test data into a statistical equation, generally on the basis of regression analyses, in order to predict a criterion. Prediction is assessed by the extent to which the sources of data can account for variance in a criterion. Depending on the nature of the study and the manner in which the data are collected, this type of research may provide evidence of either concurrent validity or predictive validity.

A recent article by Watkins and Glutting (2000) is a good example of this first approach to the study of incremental validity. Using a large sample of exceptional students and a large, nationally representative sample of students, Watkins and Glutting examined data on cognitive subtest profiles from the Wechsler Intelligence Scale for Children—Third edition (Wechsler, 1991). Specifically, Watkins and Glutting examined the incremental validity of profile characteristics (i.e., scatter and shape) over scale elevation in predicting both reading achievement and mathematics achievement. The authors conducted a series of regression analyses predicting several achievement variables; in each regression, analysis profile elevation (subtest mean score) was entered first, profile scatter (subtest standard deviation) was entered next, and profile shape (operationalized by *z* scores representing distinct clusters of overall profile shape) was entered last. Across analyses, profile elevation was found to be a statistically significant predictor of achievement scores. Adding profile scatter information to the equation did not significantly increase the prediction of achievement scores; however, the subsequent addition of profile shape data did yield a statistically significant increase in prediction.

Incremental Validity of Test-Informed Clinical Inference

A second approach to incremental validity focuses on the increment obtained from clinician-synthesized test information, not of the test scales per se. Nomothetic analyses are conducted to examine the incremental validity of idiographic judgments or interpretations made by clinicians based on a test or a series of tests in predicting a clinically relevant criterion. Given their frequent reference to decision-making tasks in clinical and personnel contexts, this conceptualization of incremental validity is consistent with the presentations by Sechrest (1963) and Wiggins (1973). Most often the research design contrasts clinical inferences based on testing with clinical inferences based on other forms of nontest data (such as unstructured clinical observations or unstructured interviews). As with the previous approach to evaluating incremental validity, depending on design factors, these types of studies may address either concurrent validity or predictive validity.

This application of incremental validity research has been infrequently studied in the last 20 years, and it should be distinguished from the Meehl-inspired (1954) research on clinical versus statistical prediction that was common from the 1950s to the early 1980s (see also Dawes, Faust, & Meehl, 1989; Grove, Zald, Lebow, Snitz, & Nelson, 2000). The latter line of research had as its focal question the superiority of clinical judgments versus statistical decision rules. In contrast, the relevant incremental validity research focuses on the extent to which test-based clinical judgments incrementally add to the validity of judgments made in the absence of test information. The following example illustrates this type of research.

Schwartz and Wiedel (1981) examined the incremental validity of judgments derived from the MMPI in neurological decision making. Six residents in either neurology or psychiatry were provided case history information, physical exam information, and medical test data (e.g., arteriograms, computed tomography scans) for thirteen patients who had been referred for neurological assessment; in half of the cases, the residents also received the patient's MMPI profile and an automated interpretation. Residents were asked to provide three possible diagnoses for each case on the basis of the information provided to them. The predicted criterion for this study was the independently derived neurological and/or psychiatric diagnoses made in the patient charts following the completion of their clinical assessments. When the residents' clinical reasoning was informed by MMPI results, diagnostic decisions were found to be significantly more accurate than when the diagnoses were made without benefit of inferences derived from the psychological test data.

Incremental Validity as Validation for New Measures

When new measures are developed or when established tests are revised, there are numerous ethical, clinical, and research factors to be considered (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). Although rarely investigated by researchers, an important consideration associated with any new

measure is its incremental validity over alternative measures available to assess the same construct. This form of incremental validity is valuable when a new test is created and when an older instrument is revised or updated (see Haynes & Lench, 2003), but it is particularly important when a new scale is created as an addition to an existing multiscale inventory. In the latter situation, it is important to justify how the new scale provides information that was formerly unavailable or less adequately obtained. Without data addressing this point, it would be possible to create an almost endless proliferation of reconfigured items or variables.

Indeed, in discussing the Minnesota Multiphasic Personality Inventory—2 (MMPI–2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) and the Minnesota Multiphasic Personality Inventory—Adolescent (MMPI–A; Butcher et al., 1992), Butcher, Graham, and Ben-Porath (1995) advocated that any new MMPI subscale or index should be evaluated to determine whether it has incremental value over existing MMPI measures. When the MMPI–2 was revised and the MMPI–A created, new items were added to the inventory, and new scales were created from these items and the original items. Therefore, when incremental validity analyses are conducted on these two tests relative to the original MMPI, the study simultaneously evaluates the added validity that comes from the revision and the newly configured scales.

As an example, Lilienfeld (1996) compared the MMPI–2 Antisocial Practices (ASP) content scale with the existing Psychopathic Deviate (Pd) scale. Using multiple samples of undergraduate students, he used several global measures of psychopathy and antisocial behavior to determine the incremental validity of each test scale relative to the other. He reported that for the majority of the criterion variables the ASP scale demonstrated significant incremental validity over the Pd scale; for several criteria, the Pd scale evidenced significant incremental validity over the ASP scale. Lilienfeld concluded, therefore, that although the two scales overlapped substantially in content, they measured different facets of the construct of psychopathy. Other investigators have found similar results when examining the full range of new MMPI–2 content scales compared with the older basic scales (e.g., Barthlow, Graham, Ben-Porath, & McNulty, 1999).

Approaching these issues from a different perspective, Goldberg (1999, in press) has articulated a very provocative challenge to test developers. His goal is to pit the relative validity of commonly used personality tests that are obtained commercially (i.e., for a fee) against the validity of parallel inventories that are freely available on the Internet. As he explained,

> For each of the constructs measured by the scales from the NEO-PI–R [revised Neuroticism–Extraversion–Openness Personality Inventory], 16PF [Sixteen Personality Factor Questionnaire], TCI [Temperament and Character Inventory], CPI [California Psychological Inventory], and HPI [Hogan Personality Inventory], parallel scales are now available in the public domain on the World-Wide-Web (http://ipip.ori.org). All of these new scales have been developed from a pool of 1,252 items, dubbed the International Personality Item Pool (IPIP, Goldberg, in press). (p. 9)

Although Goldberg discussed his challenge from the perspective of comparative validity rather than incremental validity, the comparative task exemplifies one of the critical incremental validity hurdles articulated in Sechrest's (1963) original formulation. Specifically, if the commercial inventories do not provide an increment in validity over the freely available inventories, there is no psychometric rationale for preferring them. Goldberg's initial evidence suggests that the commercial inventories do not offer superior reliability or clear incremental validity relative to their freely available counterparts. However, as with all instruments, there are factors to consider when selecting a test other than incremental validity evidence. Features that come with most commercial inventories that are not available for the Interactional Personality Item Pool counterpart scales include normative reference samples, profile forms, administration booklets, hand-scoring templates, computerized scoring options, and clinical aids to interpretation.

Design and Analysis Issues

As illustrated in these examples of incremental validity research, researchers have tended to choose correlational designs to address their incremental validity hypotheses. Indeed, with the exception of research on test-informed clinical inference, it is relatively rare to see experimental designs in this literature. In this section, we focus first on considerations relevant to experimental designs and then on issues relevant to correlational studies.

In their presentation on the treatment utility of assessment, Hayes, Nelson, and Jarrett (1987) developed a methodological typology to guide research focused on determining the contribution of assessment data to treatment assignment and outcomes (see also Finn & Tonsager, 1997, and Nelson-Gray, 2003, for discussions of similar experimental designs for treatment-relevant assessment research). If we use clinical decisions or predictions rather than treatment outcome as criteria against which the assessment data are evaluated, many of Hayes et al.'s comments and proposed research designs are directly applicable to experimental research on incremental validity. Their manipulated assessment design is especially applicable to incremental validity research, for in this design, patients are randomly assigned to two or more conditions in which the collection of assessment data or the availability of the assessment data to clinicians is varied systematically. This type of design allows the researcher to focus on the differential accuracy or validity of clinical decisions on the basis of the type and amount of assessment data available to the clinicians. Hayes et al. (1987) indicated that the manipulated assessment design could be implemented in a number of ways and could be used in both traditional between-groups designs and in time-series designs for single cases. We would add that within-subjects designs are also possible options that could be considered. A within-subjects design, for example, could involve clinicians who are asked to make predictions or diagnoses as they are given progressively more assessment data during the course of the study or as collaboration between clinician and client changes over the course of the study. The critical comparison for the purpose of incremental validity would be to determine whether the validity or accuracy of the obtained predictions improved (or deteriorated) as more data became available to the clinicians or when clients collaborated in the assessment feedback process. Alternatively, one could determine whether client ratings of symptoms or the therapeutic alliance improved as more data became available or as the test results were discussed with the client.

Both cross-sectional and longitudinal correlational designs have been used to study incremental validity issues. In most cases, researchers use multiple regression analyses to test their specific hypotheses for these types of designs. In secondary analyses in

which the original data are not available, correlations among variables can be used to calculate the incremental validity of one variable over another. According to Equation 3.3.8 from Cohen and Cohen (1983), the incremental validity of test A is a direct function of its univariate correlation with the criterion Y, test B's correlation with the criterion, and the correlation between both test A and test B such that the incremental contribution from test A is $r_A = [r_{YA} - (r_{YB}r_{AB})]/(1 - r^2_{AB})^{1/2}$. It should be noted here that the incremental *r* is also known as the semipartial *r*. Finally, as illustrated by Wiggins (1973), correlational designs that focus on classification accuracy (using categorical analyses, Bayesian statistics, or discriminant analyses) can also be used to study incremental validity.

The typical manner in which incremental validity is assessed in correlational designs is by using hierarchical multiple regression analyses to determine the contribution of one measure to the prediction of the criterion after one or more other variables have been entered into the analysis. Thus, to evaluate the incremental validity of test B in predicting a given variable, data from test A are entered into the first step of the regression analysis, and data from test B are entered on the second step. This strategy makes for a stringent test of validity because any shared variance between A and B that predicts the criterion is assigned to only A in such an analysis. Furthermore, because most researchers build regression equations using optimal variable weights rather than unit weights and because these weights are optimized in part through nonreplicable sample-dependent error, the lion's share of this sample dependent error is credited to the first variable entered into the equation, which in this case is test A.

There are a number of statistical and interpretive factors that can negatively influence the use and value of regression analyses if they are not taken into account. For present purposes, we highlight only some of these factors and we encourage readers to consult standard texts on regression analyses for further details (e.g., Aiken & West, 1991; Cohen & Cohen, 1983). As originally discussed by Sechrest (1963), a measure may have incremental validity over other measures simply because it is more reliable than other measures. This is an especially relevant factor to be considered in situations in which (a) a revised measure is compared with the previous version of the measure or (b) a newly developed measure is compared with existing measures. Depending on the context in which the measure is used, an increment in validity due to measure reliability may be either important or irrelevant. If the research has a pragmatic focus in which the goal simply is to maximize the validity or accuracy of prediction, then an increase in predictive validity due to measure reliability may be welcome. Alternatively, such an increase in predictive validity may be irrelevant if the purpose of the study is to examine whether the measure contributed truly unique variance to the prediction of the criterion. Sechrest suggested that the associations between variables could be corrected for attenuation and then used in analyses as a way to determine the extent to which a validity increment was due to differential measure reliability rather than unique information.

For both multiple regression and logistic regression models, there are a number of types of analyses that vary in the manner in which variables enter into the regression equation. To simplify our discussion, we focus specifically on the difference between stepwise and hierarchical regression approaches. In a stepwise analysis, the order of entry of variables is based on sample-specific considerations, such as the degree of association between a variable and the criterion or the degree of this association when considered in light of other variables already in the regression equation. In contrast, hierarchical analyses are ones in which the researcher specifies the order in which variables are entered into the analysis. Hierarchical approaches are preferable to stepwise approaches because a stepwise analysis tends to capitalize on sampling error and thus is likely to yield results that are not generalizable or replicable (Thompson, 1995), and a stepwise approach does not test the incremental contribution of specific measure in the same logical, theory-driven manner as in a hierarchical approach. However, it may be appropriate to use a partially stepwise procedure within a hierarchical regression analysis. For example, if a researcher is interested in controlling for the effects of a limited number of variables prior to considering the increment in predictive validity due to the measure of interest, then it can be reasonable to examine the control variables as a block and allow them to enter the equation in a stepwise fashion at the initial step of the regression analysis and then enter the measure of interest on the subsequent step of the analysis. Because the researcher is only interested in the total effect due to the initial block of variables, not the relative predictive merits of any individual variable, the analysis is logically driven by blocks of variables even though it allows for the statistical (i.e., stepwise) selection of optimal predictors within the initial block.

When using hierarchical multiple regression analysis, serious consideration must be given to the ordering of variables for entry into the equation. As we have indicated, the original presentations of incremental validity emphasized examining the predictive contributions of psychological test data beyond what can be predicted from routinely or easily collected information. What exactly constitutes easily or routinely collected data obviously varies from setting to setting. In general, though, consideration should be given to including demographic characteristics (such as age, gender, educational level, marital status, or employment status) and archival–chart data (such as prior diagnoses, school-related problems, criminal charges, or current medication use) before entering data from psychological measures in the analysis. Additionally, relatively immutable characteristics (such as gender and ethnicity) that may be causally linked to personality or psychological characteristics might also be candidates for entry into the analysis prior to the personality or psychological data (Robins, 1987).

Interpretation of how meaningful it is to have an incremental validity value of a particular size can be rather contentious (e.g., see Dawes, 1999, 2001; Perry, 2001). Once a variable has demonstrated a statistically significant increase in the prediction of a criterion beyond variables previously entered in the prediction equation, there are few guidelines to aid in determining the meaningfulness of this increase. Of course statistical significance is dependent on sample size, so that alone it cannot be used as a reliable guide, and there can be divergent views on whether a statistically significant increment has any clinical value or utility. As we indicated early in this article, Meehl (1959), in particular, emphasized the importance of evaluating this aspect of a measure or assessment procedure.

In their meta-analytic review of selection methods in personnel psychology, Schmidt and Hunter (1998) argued that the validity of a selection procedure is directly proportional to the utility of the procedure (i.e., the cost of the procedure with respect to performance of hired employees expressed either in financial terms or as

a percentage of average work output for the company). Accordingly, the incremental validity of a measure for this purpose translates directly into incremental utility, such that the percentage of increase in validity is also the percentage of increase in the utility of the test.[1] As an example, Schmidt and Hunter reported that the predictive validity of general mental ability tests with overall job performance was $r = .51$. Adding work sample tests to these ability tests yielded a combined R of .63. The increase in validity of .12 is a 24% increase in validity over what is available from using only the general mental ability tests. Thus, these authors interpreted the incremental validity value as a 24% increase in utility.

As the concept of utility in clinical contexts is somewhat different from that of the personnel selection context, this approach to evaluating the size of a validity increment may not be directly applicable to clinical assessment activities. To our knowledge, there has been no concerted attempt to produce guidelines for what might constitute a clinically meaningful validity increment. To encourage the development of such guidelines, we offer two options for consideration. First, the size of the increment could be evaluated indirectly by examining the extent to which the association between a measure and the criterion is dependent on variance shared with other variables in the regression equation. Lindenberger and Pötter (1998) termed this the shared over simple (SOS) effects approach to evaluating unique and shared effects in regression analyses. The greater the relative size of the increment in validity, the smaller the SOS effects value is. The SOS effects approach is sensitive to the amount of predictive variance the measure shares with other variables in the regression equation, and as a result, this approach may be especially applicable when the costs of assessment (including financial and human resource costs) need to be considered in determining the merit of the increment in validity. To illustrate this procedure, we consider a situation in which tests A and B are used to predict a criterion. If A accounts for 15% of the variance in the criterion (the simple effect of A), B for 20% (the simple effect of B), and A and B together for 24% (the effect when both are entered in the regression equation), the unique effect of B is 9% (i.e., 24% − 15%), and the shared effect of A and B is 11% (i.e., 20% − 9%). Thus, the SOS effects computation indicates that 55% of the variance in the criterion predicted by B is shared with A ($11/20 \times 100$). Depending on the relative costs of collecting data on tests A and B, an SOS effects value of 55% may be considered acceptable (if B is relatively inexpensive compared with A) or unacceptable (if B is relatively expensive compared with A).

The SOS effects approach does not offer an absolute metric by which to evaluate the size of the validity increment, but it could be argued that this is appropriate as the meaningfulness of the increment can only be interpreted with respect to the nature of the criterion. However, if an absolute metric is desired, we suggest a second option for evaluating the size of the validity increment that is based on the semipartial r. This statistic can be determined by obtaining the square root of the R^2 change value reported for the regression analyses, and its magnitude can be interpreted in light of some reasonable conceptual benchmarks. The first benchmark comes from Cohen's (1988, 1992) proposed guidelines for defining the magnitude of relations found in the behavioral sciences. Cohen (1988, 1992) posited that the average size of these relationships was about $r = .30$ for independently measured predictor and criterion variables. Effect size surveys from various domains of research have supported Cohen's (1988, 1992) position, with most observed relationships falling in the small to medium range (i.e., $r = .10$ to .30; see, e.g., Lipsey & Wilson, 1993; Meyer et al., 2001; Peterson, Albaum, & Beltramini, 1985). The second benchmark comes from Nunnally and Bernstein's (1994, p. 188) observation that increases in R generally are small by the time a third substantive predictive variable is included in a regression equation.

Accordingly, if we use Cohen's (1988, 1992) benchmark of $r = .30$ for an average effect, when building a typical regression equation, the first variable entered would produce $R = .30$ because $R = r$ when there is only a single variable in the equation. Next, for the second step of the equation, it would be ideal to find a semipartial r of .30 because this would mean that the new variable is of average size according to Cohen's (1988, 1992) benchmarks and also largely independent of the first variable. Considering these two variables together, the second step of a robust regression equation would produce an R of about .42 (i.e., for each variable, $r^2 = .09$; when combined, $R^2 = .18$ and thus $R = .42$). Theoretically, it would be ideal to continue adding new predictors to the equation that are largely independent of the existing predictors. If this were feasible, the third predictor (and all subsequent ones) would again produce a semipartial r of .30. However, measured variables in the social sciences are inevitably intertwined. Given this phenomenon, and consistent with Nunnally and Bernstein's (1994) observation that R increments generally are small by the third step of an equation, we propose that a semipartial r of .15 to .20 on the third step would indicate a reasonable contribution to the existing equation. Although larger than Cohen's (1988, 1992) benchmark for a small effect ($r = .10$), a contribution of this magnitude takes account of variance that is likely to be shared by all three predictor variables. A semipartial r of .15 to .20 at this step would cause R to increase slightly from .42 to between .45 and .47 (e.g., for semipartial $r = .15$, semipartial $r^2 = .0225$; when added to the R^2 of .18 from the second step of the equation, $R^2 = .2025$ and thus $R = .45$). Although others may wish to suggest alternative magnitudes for interpreting the contribution as assessed by the semipartial r, our point is that it is possible to develop general estimates of incremental validity magnitude, similar to what has been done for univariate effect sizes (e.g., Cohen, 1992).

Incremental Validity Research: The Criterion Problem

Throughout this article, we have emphasized that a typical incremental validity study is focused on ways to improve on or maximize the prediction of a criterion. Thus far, we have chosen to ignore the challenges researchers face in selecting and measuring the criteria for incremental validity research. In fact, although the criterion problem was well recognized by the time Wiggins (1973) published his influential book on personality assessment, little progress has been achieved in adequately resolving this conundrum in clinical psychology research (Garb, 1998).

Although the criterion problem is often formulated to narrowly apply just to difficulties in determining whether a decision or judgment is correct, it is useful to frame the issue in a broader

[1] As one reviewer of this article suggested, this is a controversial claim, inasmuch as it suggests that there is no need to examine other possible contributors to assessment utility once the incremental validity of a measure has been established.

manner that applies to all assessment activities. Criterion variables that have poor reliability are problematic because they produce an artificial lowering of the associations with the predictor variables, and they hamper efforts to develop valid and replicable prediction equations. Thus, whenever it is feasible to do so, researchers should attempt to improve criterion reliability or choose a more reliable criterion.

Furthermore, any increase in predictive validity that accrues simply from the association between shared systematic error in the predictor variables and the criterion (e.g., self-presentation bias that affects a predictor test and clinician ratings) is not only worthless but, in the context of clinical applications, is potentially harmful to the person who is being assessed. From a methodological perspective, a central problem is when systematic error in the criterion is aligned with the same systematic error in one of the predictors but not another. In this instance, aligned error creates artificially high associations that favor one class of predictor variables. Because systematic error is part of the true score in classical reliability theory, reliability coefficients, on their own, cannot provide an indication of the existence of this problem.

There are numerous options for improving on the criteria used for incremental validity research, most of which rely on the value of an aggregated mean or sum as a procedure for improving the reliability and validity of criterion information. When the principle of aggregation is applied to the number of items in a scale, it forms the basis of the well-known Spearman-Brown Prophecy Formula for estimating the reliability of a composite (for overviews and recent extensions see Li, Rosenthal, and Rubin, 1996, and Drewes, 2000). It has been consistently demonstrated that aggregating information over occasions (i.e., longitudinally), over stimuli (e.g., one diagnostic interview format and another), over methods of measurement (e.g., highly structured and unstructured), and over sources of information (e.g., self-report and spouse report) can enhance the reliability and validity of the aggregated information (see Epstein, 1980, 1983; Rushton, Brainerd, & Pressley, 1983). Aggregation has also been shown to be of value in improving the validity of observers' or judges' ratings (Tsujimoto, Hamilton, & Berger, 1990). The LEAD (i.e., longitudinal, expert evaluation of all data; Spitzer, 1983) approach to examining the validity of diagnostic tools also relies on aggregation, inasmuch as multiple sources and forms of data are provided to expert judges who then make diagnostic ratings on the basis of the consideration of all data available to them.

An aspect of the criterion problem that is often overlooked but that can greatly affect incremental validity results is an unrecognized or unappreciated artifact that influences the criterion variable and one (or more) but not all of the predictor scores, such that there is an artificially elevated association between the selected predictor or predictors and the criterion. The classic example of this problem in the testing literature is known as *criterion contamination,* which is defined as instances when the results from the to-be-validated test scale inform or influence the criterion designations that are used to validate the scale. For instance, if intelligence test scores are used to predict teacher ratings of intelligence, but the teacher ratings are completed after the teachers have seen the results of the intelligence test, the study would suffer from criterion contamination, and it would produce artificially high evidence of validity for the intelligence test. In an incremental validity context, the intelligence test would be artificially favored over alternative, uncontaminated predictor variables.

However, criterion contamination is just one manifestation of the underlying problem, and artifactual relations can occur in other ways. For instance, when the same source of information informs both the predictor and the criterion, the influence exerted by that source of information on both sets of variables artificially inflates estimates of their association. This can be termed a *source overlap artifact.* Methodologically, this artifact can be viewed as a variation of the well-known third variable problem in correlational research in which there seems to be an empirical association between two variables, but the association is really a function of an unmeasured third variable that influences both of the measured variables. As an example of the source overlap artifact, consider a hypothetical study in which the criterion consists of diagnoses derived from semistructured clinical interviews in a sample of clinically referred adolescents. The predictor variables for the incremental validity analysis consist of teacher ratings, parent ratings, and self-ratings. The critical issue concerns the information source that is used to obtain the criterion diagnoses. When criterion diagnoses are derived from the same source of information that generates one of the predictor variables, then that predictor and the criterion are confounded, and this should produce artificially high estimates of the predictor's validity or accuracy. Thus, if criterion diagnoses are obtained by interviewing parents, the source overlap artifact exists between the parent rating scale and the diagnostic criterion (i.e., parents know what they report in response to the test scale symptom questions and in response to the same or similar questions during the interview). The artificially elevated association between parent ratings and the diagnostic criterion also makes it more difficult for teacher ratings and adolescent self-ratings to demonstrate incremental validity. Similarly, if criterion diagnoses were obtained by interviewing the referred adolescents, the source overlap artifact would affect and artificially favor the self-report ratings relative to the parent or teacher ratings.

Another type of artifactual association can result from more purely methodological factors. For instance, consider a study that uses parent reports and self-reports of symptomatology to predict a referral status criterion that contrasts adolescents who are receiving treatment with those who are not. To the extent that adolescents are physically brought into treatment by parents who perceive a problem, regardless of how the adolescent perceives the situation, test scales derived from parental reports should have a stronger association with the criterion than those derived from the adolescents and thus receive artificial preference in an incremental validity analysis. At the same time, it is important also to note that the stronger association with parent report in this example may not all be due to an artifact of their decision-making power regarding treatment, for parents may legitimately perceive problems that adolescents fail to recognize.

Implications for Incremental Validity Research

As we indicated previously and as noted almost 20 years ago by Garb (1984), a major challenge for psychological testing and assessment incremental validity research is the noncumulative nature of much of the published research. Too frequently it seems that researchers design their studies and analyze their data without sufficient attention to how incremental validity has been assessed and analyzed in prior relevant studies. This does not mean that researchers must be unduly constrained by these earlier studies;

rather it means that, for example, there should be attempts to conceptually replicate previous findings by using similar order of entry strategies for variables in multiple regression analyses or, in experimental designs, by providing assessment data to judges in an order comparable with that found in previous research. It should also be possible in many correlational studies for researchers to explicitly conduct analyses that focus on the replication of previous results (i.e., variables are entered in the same order as was done in a previous study). In cases in which these analyses are not of focal interest for the researcher, it should be possible for the results of such analyses to be described in a few lines of text. Alternatively, researchers could ensure that a full correlation matrix of all variables is presented in their articles. As we indicated previously, there are equations using these correlations that allow for incremental validity analyses to be conducted by other interested investigators (i.e., using Equation 3.3.8 in Cohen & Cohen, 1983). Greater attention to the systematic use of either (or both) of these data reporting strategies would do much to alleviate the current difficulties facing those who wish to synthesize incremental validity findings across a research area.

Validity findings for a psychological test are always conditional, inasmuch as they are dependent on the nature of the clinical sample and criterion variable under consideration. However, incremental validity studies are doubly conditional, as any predictive variance a test shares with variables entered earlier in the regression equation is not available to be allocated to the test. As a result, efforts to replicate or generalize an incremental validity finding must include some consideration of the order of entry for variables (or the order in which assessment data are given to judges) in addition to consideration of the context of the research (e.g., formulating a diagnosis, developing a treatment plan) and the clinical sample selected to evaluate the incremental validity of a test. The doubly conditional nature of incremental validity research is another reason, in addition to those previously described, that researchers should avoid the use of stepwise regression procedures. The only instance in which stepwise entry of variables is acceptable is when the researcher is interested in controlling for the entry of a block of variables (such as demographic variables) prior to the entry of the variable of interest (such as data from a psychological test).

Finally, much incremental validity research that is intended to have direct clinical applicability focuses on assessment as a relatively static enterprise. Such research tends to rely on data collected at a single point in time that is then applied to a judgment task such as formulating a diagnosis or evaluating the outcome of an intervention. These studies do little to elucidate the incremental validity of continuous, iterative clinical assessment activities, such as the value of collecting clinical data on an ongoing basis from patients during treatment. It is relatively simple to design a study to determine whether pretreatment data from a self-report measure adds to the accuracy in predicting client diagnosis beyond what is available from other data or how much it contributes to the formulation of a clinically useful treatment plan. The situation with ongoing clinical assessment is substantially more complex, for an assessment method that may contribute little in the way of incremental validity at an initial assessment phase may prove to be important for tailoring treatment at a subsequent phase. For example, information obtained by directly observing a client who is reporting social phobic behavior may provide little incremental validity over the self-report of the client in reaching an accurate diagnosis. Such information may, however, be valuable in determining whether to target social skills deficits as part of the treatment. Much conceptual and empirical work needs to be done before the value of these clinical practices can be addressed with scientific evidence. There is some evidence, though, that assessment activities such as functional analyses have added value over other clinical data in some treatment contexts (Haynes, Leisen, & Blaine, 1997).

Implications for Clinical Assessment Practices

When conducting assessments, psychologists often focus on the importance of having convergent data that supports specific clinical conclusions and recommendations. On the one hand, to the extent that these data are derived from independent sources of information and share minimal method variance, there is certainly value in obtaining convergent data that supports the same clinical conclusion such as a diagnosis or a recommendation for a specific type of psychological service. On the other hand, consistent with our discussion of artifacts, if the data sources are based on what is essentially the same source or form of information, then the apparently convergent data provide little more than an unwarranted sense of security regarding the validity or accuracy of the conclusions. Moreover, as Tsujimoto et al. (1990) have demonstrated, the accuracy of predictions increases when data sources provide nonredundant information. Accordingly, the desire to seek confirmatory evidence in clinical assessments needs to be balanced with the recognition that nonoverlapping sources of data are required for improving the accuracy of clinical decisions.

In our discussion of research conceptualizations of incremental validity, we distinguished between the incremental validity of testing measures and the incremental validity of test-informed clinical inferences. This distinction is extremely important in attempts to use research findings in an applied context, for even if a study demonstrates incremental validity for a measure, this does not provide evidence that, in practice, psychologists using the measure show improved prediction of the criterion. Regression analyses combine data in an optimal manner, thus leading to an overall reduction of prediction errors if the prediction rule is used in a clinical setting. As has been known for quite some time, in comparison with statistically derived prediction rules, people are less accurate in consistently combining test data (e.g., Dawes et al., 1989; Grove et al., 2000; Meehl, 1954). Accordingly, evidence for the incremental validity of a measure does not necessarily translate directly into improvements in everyday assessment decisions in which the measure is used.

Perhaps the biggest challenge facing those who wish to base their clinical assessment services on incremental validity evidence is that, assuming there was available a cumulative literature on which to draw, there is no user-friendly strategy or procedure currently in place to facilitate the application of nomothetic, data-based findings to the individual clinical case. As we indicated previously, even seminal writings on incremental validity have not addressed how incremental validity evidence might be applied to a single case. Although a number of authors have suggested that signal detection theory and receiver operating characteristics curves can provide avenues for translating validation data into user-friendly decision-making tools (e.g., McFall & Treat, 1999; Swets, Dawes, & Monahan, 2000), there has been relatively little progress to date on this front. Of course, incremental validity

research on test-informed clinical inference provides directly applicable findings for clinical assessment, as the analyses specifically examine the incremental validity of idiographic judgments or interpretations made by clinicians, based on test data, in predicting clinical criteria. Unfortunately, compared with the range of purposes for which clinicians conduct assessments, the scope of this literature is relatively limited and, therefore, can not yet provide sufficient empirically based guidance for commonly encountered assessment tasks. As a result, additional research on the utility of all forms of psychological test data for various commonly encountered clinical assessment situations is sorely needed (Hunsley & Bailey, 2001; Meyer & Archer, 2001).

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

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. *Psychological Bulletin, 51,* 201–238.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Barthlow, D. L., Graham, J. R., Ben-Porath, Y. S., & McNulty, J. L. (1999). Incremental validity of the MMPI–2 content scales in an outpatient mental health setting. *Psychological Assessment, 11,* 39–47.

Boland, A., & Cappeliez, P. (1997). Optimism and neuroticism as predictors of coping and adaptation in older women. *Personality & Individual Differences, 22,* 909–919.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory 2: Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Butcher, J. N., Graham, J. R., & Ben-Porath, Y. S. (1995). Methodological problems and issues in MMPI, MMPI–2, and MMPI–A research. *Psychological Assessment, 7,* 320–329.

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S., & Kaemmer, B. (1992). *MMPI–A (Minnesota Multiphasic Personality Inventory—Adolescent).* Minneapolis: University of Minnesota Press.

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist, 15,* 546–553.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Conway, J. M., Lombardo, K., & Sanders, K. C. (2001). A meta-analysis of incremental validity and nomological networks for subordinate and peer ratings. *Human Performance, 14,* 267–303.

Cronbach, L. J., & Gleser, G. C. (1957). *Psychological tests and personnel decisions.* Urbana: University of Illinois Press.

Dawes, R. M. (1999). Two methods for studying the incremental validity of a Rorschach variable. *Psychological Assessment, 11,* 297–302.

Dawes, R. M. (2001). Incremental validity of the Ego Impairment Index: It's fine when it's there. *Psychological Assessment, 13,* 408–409.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989, March). Clinical versus actuarial judgment. *Science, 243,* 1668–1674.

Donnay, D. A. C., & Borgen, F. H. (1999). The incremental validity of vocational self-efficacy: An examination of interest, self-efficacy, and occupation. *Journal of Counseling Psychology, 46,* 432–447.

Drewes, D. W. (2000). Beyond Spearman-Brown: A structural approach to maximal reliability. *Psychological Methods, 5,* 214–227.

Elliott, A. N., O'Donohue, W. T. O., & Nickerson, M. A. (1993). The use of sexually anatomically detailed dolls in the assessment of sexual abuse. *Clinical Psychology Review, 13,* 207–221.

Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist, 35,* 790–806.

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality, 51,* 360–392.

Finn, S. E., & Tonsager, M. E. (1997). Information-gathering and therapeutic models of assessment: Complementary paradigms. *Psychological Assessment, 9,* 374–385.

Garb, H. N. (1984). The incremental validity of information used in personality assessment. *Clinical Psychology Review, 4,* 641–655.

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

Glutting, J. J., Youngstrom, E. A., Ward, T., & Ward, S. (1997). Incremental efficacy of WISC–III factor scores in predicting achievement: What do they tell us? *Psychological Assessment, 9,* 295–301.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Goldberg, L. R. (in press). The comparative validity of adult personality inventories: Applications of a consumer-testing framework. In S. R. Briggs, J. M. Cheek, & E. M. Donahue (Eds.), *Handbook of adult personality inventories.* New York: Plenum Press.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19–30.

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory.* Minneapolis: University of Minnesota Press.

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., Leisen, M. B., & Blaine, D. D. (1997). Design of individualized behavioral treatment programs using functional analytic clinical case methods. *Psychological Assessment, 9,* 334–348.

Haynes, S. N., & Lench, H. C. (2003). Incremental validity of new clinical assessment measures. *Psychological Assessment, 15,* 456–466.

Haynes, S. N., & O'Brien, W. O. (2000). *Principles of behavioral assessment: A functional approach to psychological assessment.* New York: Plenum Press.

Holden, R. R., & Hibbs, N. (1995). Incremental validity of response latencies for detecting fakers on a personality test. *Journal of Research in Personality, 29,* 362–372.

Huffcutt, A. I., Roth, P. L., & McDaniel, M. A. (1996). A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology, 81,* 459–473.

Hunsley, J. (1987). Cognitive processes in mathematics anxiety and test anxiety: The role of appraisals, internal dialogue, and attributions. *Journal of Educational Psychology, 79,* 388–392.

Hunsley, J., & Bailey, J. M. (2001). Whither the Rorschach? An analysis of the evidence. *Psychological Assessment, 13,* 472–485.

James, S., & Hunsley, J. (1995). The Marital Adaptability and Cohesion Evaluation Scale III: Is the relation with marital adjustment linear or curvilinear? *Journal of Family Psychology, 9,* 458–462.

Johnston, C., & Murray, C. (2003). Incremental validity in the psycholog-

ical assessment of children and adolescents. *Psychological Assessment, 15,* 496–507.

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2002). Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct? *Journal of Personality and Social Psychology, 83,* 693–710.

Kline, R. B., Snyder, J., Guilmette, S., & Castellanos, M. (1992). Relative usefulness of elevation, variability, and shape information from WISC–R, K-ABC, and fourth edition Stanford-Binet profiles in predicting achievement. *Psychological Assessment, 4,* 426–432.

Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods, 1,* 98–107.

Lilienfeld, S. O. (1996). The MMPI–2 Antisocial Practices content scale: Construct validity and comparison with the Psychopathic Deviate scale. *Psychological Assessment, 8,* 281–293.

Lilienfeld, S. O. (1997). The relation of anxiety sensitivity to higher and lower order personality dimensions: Implications for the etiology of panic attacks. *Journal of Abnormal Psychology, 106,* 539–544.

Lindenberger, U., & Pötter, U. (1998). The complex nature of unique and shared effects in hierarchical linear regression: Implications for developmental psychology. *Psychological Methods, 3,* 218–230.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181–1209.

Lofland, K. R., Cassisi, J. E., Levin, J. B., Palumbo, N. L., & Blonsky, E. R. (2000). The incremental validity of lumbar surface EMG, behavioral observation, and a symptom checklist in the assessment of patients with chronic low-back pain. *Applied Psychophysiology & Biofeedback, 25,* 67–78.

Mash, E. J., & Hunsley, J. (2004). Behavioral assessment: Sometimes you get what you need. In S. N. Haynes & E. M. Heiby (Eds.), *The comprehensive handbook of psychological assessment: Vol. 3. Behavioral assessment* (pp. 489–501). New York: Wiley.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50,* 215–241.

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

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

Meyer, G. J. (2000). Incremental validity of the Rorschach Prognostic Rating Scale over the MMPI Ego Strength Scale and IQ. *Journal of Personality Assessment, 74,* 356–370.

Meyer, G. J., & Archer, R. P. (2001). The hard science of Rorschach research: What do we know and where do we go? *Psychological Assessment, 13,* 486–502.

Meyer, G. J., Finn, S. E., Eyde, L., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128–165.

Murray, H. A. (1943). *Thematic Apperception Test manual.* Cambridge, MA: Harvard University Press.

Nelson-Gray, R. O. (2003). Treatment utility of psychological assessment. *Psychological Assessment, 15,* 521–531.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Perry, W. (2001). Incremental validity of the Ego Impairment Index: A re-examination of Dawes (1999). *Psychological Assessment, 13,* 403–407.

Peterson, R. A., Albaum, G., & Beltramini, R. F. (1985). A meta-analysis of effect sizes in consumer behavior experiments. *Journal of Consumer Research, 12,* 97–103.

Robins, C. J. (1987). On interpreting results of multiple regression procedures: A cautionary note for researchers and reviewers. *Cognitive Therapy and Research, 11,* 705–708.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94,* 18–38.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Schwartz, C. E., Kozora, E., & Zeng, Q. (1996). Toward patient collaboration in cognitive assessment: Specificity, sensitivity, and incremental validity of self-report. *Annals of Behavioral Medicine, 18,* 177–184.

Schwartz, S., & Wiedel, T. C. (1981). Incremental validity of the MMPI in neurological decision-making. *Journal of Personality Assessment, 45,* 424–426.

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

Siem, F. M. (1996). The use of response latencies to enhance self-report personality measures. *Military Psychology, 8,* 15–27.

Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry, 24,* 399–411.

Stout, C. E., & Cook, L. P. (1999). New areas for psychological assessment in general health care settings: What to do today to prepare for tomorrow. *Journal of Clinical Psychology, 55,* 797–812.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1–26.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55,* 525–534.

Tsujimoto, R. N., Hamilton, M., & Berger, D. E. (1990). Averaging multiple judges to improve validity: Aid to planning cost-effective clinical research. *Psychological Assessment, 2,* 432–437.

Watkins, M. W., & Glutting, J. J. (2000). Incremental validity of WISC–III profile elevation, scatter, and shape information for predicting reading and math achievement. *Psychological Assessment, 12,* 402–408.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children—Third edition.* San Antonio, TX: The Psychological Corporation.

Widiger, T. A., & Schilling, K. M. (1980). Toward a construct validation of the Rorschach. *Journal of Personality Assessment, 44,* 450–459.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment.* Menlo Park, CA: Addison Wesley.

Woodard, J. L., & Axelrod, B. N. (1995). Parsimonious prediction of Wechsler Memory Scale—Revised memory indices. *Psychological Assessment, 7,* 445–449.

Yates, B. T., & Taub, J. (2003). Assessing the costs, benefits, cost-effectiveness, and cost-benefit of psychological assessment: We should, we can, and here's how. *Psychological Assessment, 15,* 478–495.

Received December 23, 2002
Revision received July 3, 2003
Accepted July 8, 2003 ■