Incremental Validity of New Clinical Assessment Measures

Stephen N. Haynes
University of Hawaii at Manoa

Heather C. Lench
University of California, Irvine

The authors address conceptual and methodological foundations of incremental validity in the evaluation of newly developed clinical assessment measures. Incremental validity is defined as the degree to which a measure explains or predicts a phenomenon of interest, relative to other measures. Incremental validity can be evaluated on several dimensions, such as sensitivity to change, diagnostic efficacy, content validity, treatment design and outcome, and convergent validity. Indices of incremental validity can vary depending on the criterion measures, comparison measures, and individual differences in samples. The authors review the rationale for, principles, and methods of incremental validation, including the selection of comparison and criterion measures, and address data analytic strategies and the conditional nature of incremental validity evaluations in the selection of measures. Incremental validity contributes to, but is different from, cost–benefits, which reflect the cost of acquiring the data and the benefits from the data. The impact of an incremental validity index on whether a measure is selected will be moderated by the cost of acquiring the new data, the importance of the measured phenomenon, and the clinical utility of the new data.

From 1998 through the first half of 2002, 298 manuscripts submitted to *Psychological Assessment* reported on the development and validation of a new assessment instrument or scale.[1] Only 26 of those manuscripts addressed the relative or incremental validity of the new instrument—that is, the degree to which the instrument provided measures that were more valid than alternative measures of the same variables. Eleven manuscripts presented new instruments to measure depression, 10 manuscripts presented new instruments to measure alcohol and drug use, and 9 manuscripts presented new instruments to measure child behavior problems—three areas in which there are existing assessment instruments that have undergone psychometric evaluation. One consequence is that clinical and research assessors must select from an array of instruments designed to measure the same phenomenon without information on their relative efficacy.

Infrequent attention to the incremental contributions of new assessment instruments persists despite several decades of advocacy by assessment scholars.[2] In 1963, Sechrest published an article that is often cited as the first to emphasize the importance of incremental validity evaluations with all new clinical assessment instruments. However, earlier studies had also discussed the relative validity of multiple measures used for the same assessment purpose (Grant, 1938; Sangren, 1929; Tupes, 1950).

The degree to which a new measure contributes to clinical judgments has continued to be a topic of scholarly interest (Barthlow, Graham, Ben-Porath, & McNulty, 1999; Dawes, 1999, 2001; Garb, Wood, Nezworski, Grove, & Stejskal, 2001; Haynes & O'Brien, 2000). Two recent special sections in *Psychological Assessment* (1999, 2001), edited by Greg Meyer, included several articles that addressed the degree to which measures from the Rorschach could contribute to clinical judgments above the contributions by other standardized questionnaires and interviews.

Incremental validity supplements traditional psychometric dimensions of content, convergent, predictive, discriminant, and other forms of validity (see discussions in Foster & Cone, 1995; Haynes, Nelson, & Blaine, 1999; Nunnally & Bernstein, 1994; Silva, 1993), because it addresses the performance of a measure relative to others. Given that a measure has been shown to validly predict a phenomenon of interest, does the new measure (a) predict the phenomenon more validly or accurately than other measures, (b) contribute meaningfully to predictive efficacy when added to already-existing or more readily obtainable measures, and (c) cost less than other measures? Thus, incremental validation is a

---

Correspondence concerning this article should be addressed to Stephen N. Haynes, Department of Psychology, University of Hawaii at Manoa, 2430 Campus Road, Honolulu, Hawaii 96822, or to Heather C. Lench, Department of Psychology and Social Ecology, University of California, Irvine, 3340 Social Ecology I, Irvine, California 92697. E-mail: sneil@hawaii.edu or xiola96@hotmail.com

[1] This is out of approximately 1,800 submitted manuscripts submitted during this period. Each manuscript submitted to *Psychological Assessment* was coded on approximately 150 variables, in domains of assessment method, sample composition, focus, and type of psychometric analyses.

[2] Many articles discuss the incremental validity of assessment *instruments.* More precisely, we are interested in the incremental validity of *measures,* or scores derived from instruments. Many assessment instruments provide multiple measures that can differ in validity and utility. Furthermore, as we discuss later, the validity and utility of a measure can vary across populations, assessment goals (the judgments affected by the measure), and the criterion measures used to establish its validity.

complex and demanding addition to standard psychometric evaluations.

In this article, we focus on the incremental validity of new clinical assessment measures—the incremental validity of new measures from an existing assessment instrument, from the refinement of an existing instrument, or from a new assessment instrument. We consider conceptual and methodological issues in evaluating the degree to which these new measures perform better or more cost-effectively than existing measures. We also offer recommendations for methods of incremental validation and discuss factors that affect the meaning of incremental validity indices.

Several applications of incremental validity concepts are not covered in this article. This article does not address the selection of items or behavior observation codes in the development of an instrument, the incremental validity of adding more time-sampling intervals or situations in observation assessment (see discussion in Haynes & O'Brien, 2000, and Smith, Fisher, & Fister, 2003). We also do not address the incremental validity of adding a measure to an existing multimethod assessment strategy or of adding an additional measure (e.g., using multiple informants rather than one informant) to a monomethod assessment strategy. Garb (2003) discusses in this section the comparative and incremental validity of interviews, personality questionnaires, projective methods, and brief self-report instruments in the measurement of adult psychopathology. Other discussions and examples of these strategies can be found in Garb (1985); Lofland, Cassisi, Levin, Palumbo, and Blonsky (2000); and Power, Andrews, et al. (1998). However, the hierarchical linear regression strategies discussed later in this article are amenable to the investigation of incremental validity of multiple methods and multiple instruments.

In addressing incremental validity in the evaluation of new measures, we first consider the definitions, domain, and facets of incremental validity. We then consider the assessment contexts in which the incremental validity evaluations are warranted. Next, we discuss principles and methods of incremental validation. Finally, we discuss caveats and the conditional nature of incremental validity.

Definition of Incremental Validity

Definitions of *incremental validity* have varied across assessment scholars (see definitions of incremental validity in Barnett, Lentz, & Macmann, 2000; Cronbach & Gleser, 1957; Dawes, 1999; Elliott, O'Donohue, & Nickerson, 1993; Foster & Cone, 1995; Garb, 1985; Haynes & O'Brien, 2000; Murphy-Berman, 1994; Schwartz & Wiedel, 1981; Sechrest, 1963; Wiggins & Kohen, 1971). These definitions differ in the degree to which they incorporate concepts of relative cost-effectiveness, additional versus comparative validity, and relative validity compared with one versus multiple instruments and statistical versus clinically based judgments of incremental validity. However, they have in common the idea of relative predictive efficacy: that a measure has incremental validity to the degree to which it increases the ability to predict an important phenomenon. Congruent with the main tenets of past definitions, the definition of *incremental validity* we have adopted is: "The degree to which a measure explains or predicts some phenomena of interest, relative to other measures."

The Dimensions and Conditional Nature of Incremental Validity

There are several implications of our definition of incremental validity. First, there can be *multiple dimensions of incremental validity;* that is, incremental validity can be measured in many ways, depending on the clinical judgments that are to be made or the goals of assessment. For example, in evaluating the incremental validity of a new self-report measure of bipolar manic episodes, we may be interested in several dimensions: (a) the degree to which the new measure is more sensitive to changes in symptoms (incremental sensitivity to change); (b) the degree to which the new measure more adequately captures the range of manic behaviors outlined in the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM–IV*; American Psychiatric Association, 1994; incremental content validity); (c) the degree to which the new measure is more strongly correlated with informant reports of respondents' manic behaviors (incremental convergent validity) or less strongly correlated with other behavior disorders (incremental discriminant validity); and (d) the degree to which the new measure identifies functional relations and controlling events for manic behaviors, which can guide treatment decisions (incremental treatment utility). As this example illustrates, estimates of the incremental validity of a new measure can vary depending on which dimension of evaluation is being considered. Table 1 outlines several dimensions upon which incremental validity can be evaluated.

Second, inferences about the incremental validity of a new measure will depend on which criterion measures are being explained or predicted. For example, estimates of the ecological incremental validity (see Table 1) of a new analogue behavioral observation measure of marital problem-solving behaviors (see review in Heyman, 2001) will depend on which assessment method is used—for example, whether self-report, informant report, self-monitoring, or external observer methods are used to estimate problem solving at home. Furthermore, incremental validity inferences will depend on which assessment instruments, within a method, and which measures from those instruments are used.

Third, inferences about the incremental validity of a new measure will depend on the alternative measures with which it is compared. For example, estimates of the incremental validity of a new measure of adult depression will vary, depending on whether the comparison measure is the Beck Depression Inventory (BDI), the Center for Epidemiological Studies—Depression Scale, the Minnesota Multiphasic Personality Inventory (MMPI) Depression scale, or another measure of depression (see review in Dozois & Dobson, 2002).

Fourth, inferences about incremental validity apply to measures from an assessment instrument, not the instrument itself. Most assessment instruments can provide multiple measures, which can differ in their degree of incremental validity. Some measures may be based on aggregation of all items or behavior codes in the instrument. Other measures from the same instrument may be composed of varying combinations of items and codes, to form subscales.

Fifth, inferences about the incremental validity of a new measure can vary across target populations and samples. Incremental validity indices can be affected by the sex, age, ethnicity, socio-

Table 1
Dimensions of Incremental Validity

Dimension	Definition	Example
Incremental positive and negative predictive power	Degree to which a measure increases the proportion of individuals identified by an instrument as having (or not having) a disorder, or emitting or not emitting a behavior, who truly have (or do not have) the disorder or who truly do or do not emit a behavior; diagnostic efficacy.	Inconclusive evidence of the degree to which MMPI–2 supplementary scales detected respondents who underreported symptoms better than traditional L and K subscales (Baer & Miller, 2002).
Incremental content validity	Degree to which the elements of an assessment instrument (e.g., behavior codes, items) tap a broader range of facets of a targeted construct.	Some models and subscales of the WAIS defined the construct and covered the facets of intelligence better than other WAIS models and subscales (Bowden, Carstairs, & Shores, 1999).
Incremental treatment design validity	Degree to which a measure leads to the design of more effective treatments for a client.	The Ego Impairment Index of the Rorschach, compared to self-report, was a better predictor of the degree to which antidepressant treatment would be effective (Perry & Viglione, 1991).
Incremental sensitivity to change	Degree to which a measure is more sensitive to changes in a variable.	The Sickness Impact Profile was more sensitive to clinical changes in physical and psychological functioning during hospitalization than were measures from four other instruments (J. N. Katz, Larson, Phillips, & Fossel, 1992).[a]
Incremental predictive validity	Degree to which a measure can account for a higher proportion of variance in a variable measured later.	Subscale scores on the WISC–III increased the predictive efficacy of academic achievement as measured by the WIAT, above that of the Full Scale score, was statistically significant but clinically insignificant (Glutting et al., 1997).
Incremental criterion validity	Degree to which a measure accounts for a higher proportion of variance in a criterion measure.	WISC–III elevation, but not scatter and shape, accounted for variance in academic achievement, as measured by the WIAT and WJ–R (reading and math), among exceptional and nonexceptional students (Watkins & Glutting, 2000).
Incremental convergent or discriminant validity	Degree to which a measure accounts for a higher proportion of variance in similar measures of related constructs or does not exhibit high degrees of shared variance with measures of dissimilar constructs.	MMPI content scales accounted for a significantly greater proportion of the variance than MMPI clinical scales in predicting symptoms in outpatient adults (7/10 for males and 3/10 for females; Barthlow et al., 1999).
Incremental discriminative validity (similar to positive and negative predictive power)	Degree to which a measure accurately identifies persons placed into groups on the basis of another measure.	The Ego Impairment Index of the Rorschach accounted for a significantly greater proportion of the variance than the index component scales of the index in several positive symptoms (e.g., hallucinations, delusions) of schizophrenic individuals (Perry, 2001).
Ecological validity or generalizability across settings	Degree to which an assessment instrument provides measures that are indicative of the behavior of a person in the natural environment.	Analogue Behavioral Observation was more effective than some self-report measures in assessing some aspects of marital conflict (Heyman & Slep, 2003).

Note. MMPI–2 = Minnesota Multiphasic Personality Inventory—2; WAIS = Wechsler Adult Intelligence Scale; WISC–III = Wechsler Intelligence Scale for Children—III; WIAT = Wechsler Individual Achievement Test; WJ–R = Woodcock-Johnson Psycho-Educational Battery–Revised.
[a] Other measures were derived from the Medical Outcomes Study (MOS) 36-Item Short-Form Health Survey, the Functional Status Questionnaire, the Modified Health Assessment Questionnaire, and a shortened version of the Arthritis Impact Measurement Scales.

economic status, cognitive or physical abilities, intensity of the measured phenomenon, or other dimension of individual differences. For example, a new analogue behavioral observation measure of marital problem-solving behaviors may be better than extant measures for couples in outpatient treatment but not for couples in which one spouse is in a psychiatric inpatient unit. As Garb (2003) discussed, the incremental validity of the posttrau-

matic stress disorder (PTSD) measure from the MMPI may differ as a function of the emotional state of the respondent when he or she is completing the instrument.

These five aspects of incremental validity emphasize its conditional nature: Inferences about the incremental validity of a new measure can vary as a function of the goals of assessment, the criterion and comparison measures selected, and the sample with

which it is used. As we discuss later, these conditional aspects have implications for methods of incremental validation. To select the most appropriate dimensions of incremental validation, and to guide the incremental validation process, developers of a new measure must (a) decide how the new measure is to be used (the clinical judgments that will be based on it), (b) select the criteria on which validity inferences are to be based, (c) select the alternative measures with which the new measure will be compared, and (d) identify the population with which it is to be used and select a representative sample of that population.

Incremental Validity and Cost-Effectiveness Analyses

Our definition of incremental validity does not reflect cost-effectiveness considerations. Cost–benefit reflects the cost (e.g., financial, time) of acquiring the assessment data and benefits of the data for clients (e.g., improved treatment outcome). Yates and Taub (2003) differentiated cost–benefit analysis (which addresses the costs of an assessment to its monetary outcomes, or benefits) and cost-effectiveness analysis (which addresses whether a given measure is better than another, despite additional costs). Clinical utility is an even higher order and more qualitative dimension. It incorporates validity, cost-effectiveness, applicability and practicality, transportability across assessors and settings, acceptability of the assessment procedure to assessors and respondents, and the consequential validity of judgments based on the assessment data. Thus, cost-effectiveness and clinical utility evaluations are affected by, but are not components of, incremental validity (see discussion of cost-effectiveness in Hargreaves, Shumway, Hu, & Cuffel, 1998; Heinrichs, 1990; Yates & Taub, 2003; see discussion of consequential validity in Messick, 1994).

Incremental cost-effectiveness considerations usually are important after a measure has been shown to have incremental validity in comparison to other measures. After the incremental validity of a new measure has been supported, two questions are addressed: Given that a new measure provides data that are more valid than those provided by other measures, (a) what is the relative cost of acquiring these data, compared with data from the comparison measures, and (b) is it cost-effective—does the ratio of costs to benefits of the new measure, relative to that of others, warrant its use?

A cost-effectiveness analysis can support the use of a new measure that is less valid but much less costly than alternative measures. The outcomes of cost-effectiveness analyses are affected by the importance of the clinical judgments that are being made. Relative cost will be less important than incremental validity with more important behaviors and judgments. For example, one may be less concerned with the cost associated with a better measure if it aids in making judgments about the care of children at risk for abuse or neglect or if it identifies variables associated with recidivism of psychotic symptoms or substance dependence (e.g., Murphy-Berman, 1994) rather than with the cost associated with judgments about less severe behavior problems.

A final note on the conditional nature of costs-effectiveness is that, congruent with the multiple dimensions of incremental validity noted in Table 1, the costs and benefits of a new measure can also be evaluated on multiple dimensions. Costs and benefits are often measured on financial dimensions, as Yates and Taub (2003) suggested. For example, do the new data help design a treatment

that speeds the client's return to work? What is the financial cost of acquiring the data (e.g., therapist time, costs associated with scoring and interpreting test results)? Costs and benefits can be measured on health dimensions. Do the new data help in selecting treatments to reduce medication use and to reduce functional impairment associated with chronic diseases? Costs and benefits can be measured in regard to role functioning (e.g., performance in work and parental roles), treatment durations, consequences for others (e.g., effects on the children and spouses of clients), quality of life (e.g., life satisfaction and enjoyment), and many other dimensions.

The costs–benefits of a new measure can also depend on the characteristics of the assessment targets. For example, it may not be cost-beneficial to administer a complex set of assessment instruments to construct a case formulation for all clients seeking treatment at an outpatient center; however, it may be cost-beneficial to do so for clients who are not progressing well in therapy.

In sum, although we focus in this article on the incremental validity of a new measure, cost-effectiveness is also an important dimension. Yates and Taub (2003) discuss cost-effectiveness in greater depth in their article in this special section. In the Discussion section, we discuss additional factors that affect the ultimate decisions about the applicability and utility of a new measure.

When Are the Development and Incremental Validation of a New Measure Warranted?

Table 1 suggests several conditions in which the development of a new and incrementally valid measure would be warranted. Unsatisfactory performance of existing measures on any of the evaluative dimensions listed in Table 1 would suggest the need for a new instrument, the refinement of existing instruments, or the derivation of new measures (i.e., new scales) from an existing instrument. As Smith and McCarthy (1995) discussed, unsatisfactory performance across studies on any of these ultimate outcomes (e.g., positive predictive power, sensitivity to change) can depend on lower level difficulties with an instrument. These difficulties may include problems with item construction and content, the hierarchical structure of an instrument, and internal consistency within scales. Smith et al.'s (2003) contribution to this special section considers in greater detail item-level problems that can limit the validity of a measure and strategies for instrument construction to maximize incremental validity.

A new and incrementally valid measure might also be warranted when a measure performs differentially across dimensions of individual difference. For example, a measure of intimate-partner aggression may perform well for adults but may not tap important aspects of aggression that occurs among dating adolescents—that is, it may have poor content validity for adolescents.

When confronted with poorly performing measures, how does one know if the best strategy is to refine or extract a new measure from an existing instrument or to develop a new one? Several data analytic strategies can inform this decision. For example, item-response methods can help detect item bias or items that are otherwise performing unsatisfactorily (Embretson & Prenovost, 1999). Internal consistency, item-level temporal stability evaluations, interobserver agreement indices for individual behavior codes, item–factor loadings in factor analysis, item–total correla-

tion, item dispersions, and item–criterion correlation analyses can also help detect poorly performing items.

The decision to refine an instrument or develop a new one depends in part on the proportion of items that are performing poorly. A few poorly performing items can be modified or deleted; however, deletion is an option only if coverage of the domain and facets of the targeted variable would not be compromised. For example, if poorly performing items were all focused on an avoidance facet of PTSD, their deletion would result in higher validity indices on many dimensions but a concomitant reduction in the content validity of the instrument.

The development of a new measure also can be warranted even when items and measures from an existing instrument perform well (e.g., Strom, Gray, Dean, & Fischer, 1987). This scenario is possible when measures from an instrument cover some but not all facets of a multifaceted construct. Consider an instrument for measuring PTSD that includes valid measures of most facets of PTSD, but not the avoidance facet. The measure could show high levels of validity on many dimensions listed in Table 1, but not content validity (Haynes, Richard, & Kubany, 1995), because the resulting measures were not representative of the domain of the PTSD construct, as suggested by *DSM–IV*. Improved content validity should be reflected in greater predictive efficacy. The new measures could have greater convergent validity. In summary, the construction of a new measure from scratch is suggested when item-level analyses identify many poorly performing items or a dearth of items that tap important facets of the targeted construct.

The failure of an instrument to cover all facets of a multifaceted construct, the representativeness of items (one aspect of content validity, Haynes et al., 1995), is a forceful rationale for the development of a new measure (Strom et al., 1987). For instruments with inadequate coverage of a construct, revision of existing items would be insufficient to ensure coverage of all facets, and new items would have to be constructed.

The content validity of an assessment instrument, along with the discriminant validity of its measures, can also be enhanced by the omission of irrelevant items, that is, items that tap constructs other than the targeted construct. The content and discriminant validity of an instrument that contains items that tap constructs outside of the targeted domain (e.g., a depression measure that includes items that also measure anxiety) are compromised and can be strengthened with the omission of these items.

As we noted earlier, the development of a new measure can also be warranted if a measure is valid on a dimension of interest but is costly. For example, there are several self-report instruments for measuring depressive behaviors validated in clinical assessment settings (Dozois & Dobson, 2002). However, most of these involve 15–30 items, which may be too costly for use in community-based epidemiological studies when the goal is to provide an accurate but cost-effective index of multiple behavior problems.

The development of a new measure does not guarantee that it will be more valid than other measures. A new measure may be significantly associated with an important outcome variable yet perform worse than other measures of the same outcome (Sechrest, 1963). Reflecting the conditional nature of validation, the new measure may perform better than others on some dimensions, in some settings, or for some populations, but not others. In the following sections we discuss methods of evaluating the incremental validity of a new measure.

Considerations in the Development and Incremental Validation of a New Measure

Principles of Instrument and Measure Development

Methods for developing a new assessment instrument and measures to maximize its validity and utility have been extensively discussed (e.g., Foster & Cone, 1995; Hambleton & Zaal, 1991; Nunnally & Bernstein, 1994; Smith & McCarthy, 1995) and are outside the domain of this article. However, we emphasize that several aspects of the early development phase (e.g., the construction, refinement, and selection of items and behavior codes) strongly affect the content and, ultimately, the incremental validity of the resulting measure. To increase the chance that a new measure will contain items that are relevant to and representative of the targeted construct and will perform better than extant measures, developers should (a) specify the goals of assessment, that is, how the measures will be used (e.g., treatment planning, screening, or diagnosis); (b) define the population with which the measures will be used (e.g., older vs. younger adults, persons in the community vs. persons in a psychiatric hospital); (c) specify the targeted construct, its domain and facets, to ensure that they are adequately covered by the items; (d) examine the empirical and nonempirical literature on the strengths and weaknesses of alternate measures; (e) construct items for the new instrument that are congruent with items a–d (based on suggestions from the empirical literature about the targeted construct and population, an examination of item performance in extant instruments, and population sampling); and (f) follow standard psychometric procedures for evaluation of item performance, internal structure, internal consistency, temporal stability, accuracy, and construct validity.

Selection of Criteria for Use in Incremental Validation

Earlier in this article we noted that incremental validity can be evaluated on multiple dimensions. Further complicating the validation process is the fact that many criteria can be used for validation on each dimension. As Blais, Hilsenroth, Castlebury, Fowler, and Baity (2001) noted, the purpose of all validation is to predict an important phenomenon, and the criteria selected to measure the phenomenon will affect incremental-validity inferences. Sangren noted this in 1929 when he found differences in the relative validity of seven tests of intelligence with first-grade children, depending on which criterion for intelligence was used.

The measurement of PTSD symptoms illustrates the importance of criterion measures. Indices of incremental validity for a new self-report measure of PTSD would be expected to differ depending on whether it was used to predict treatment outcome, monitor changes during treatment or across time, identify symptoms or functional relations to aid treatment planning, predict impairment in daily activities, predict health outcomes, or show agreement with diagnosis based on structured diagnostic interviews. Furthermore, there are multiple ways to arrive at a diagnosis, multiple measures of health outcomes, and multiple ways to estimate treatment outcome. Differential-validity indices across criteria reemphasize our earlier point that the measure's developer must clearly specify the goals of the assessment and the purposes for which a measure will be used before constructing items for a new assessment instrument.

The effect of different criteria on incremental validity indices is illustrated by Blais et al.'s (2001) study. Using hierarchical regression analyses, the authors found that both MMPI–2 and Rorschach measures added incrementally to the prediction of *DSM–IV* borderline and narcissistic personality disorder criteria but not to the prediction of *DSM–IV* histrionic and antisocial personality disorder criteria.

Criteria for validation should be based on direct, minimally inferential measures of clinically important phenomena. For the new self-report questionnaire measure of PTSD, covariance with measures of important clinical phenomena from other methods (e.g., interviewer report, self-monitoring, biological markers, external observers, medical records) would allow for stronger inferences about its incremental validity than would covariance with another aggregated self-report measure of PTSD symptoms. Aggregated self-report measures are often a less direct, although necessary, measure of many PTSD symptoms. With two measures based on the same method, the magnitude of covariance can partially reflect item contamination (i.e., when two measures contain similar items), shared method variance, and test–retest reliability.

Selection of the Comparison Measures

After a new measure has been constructed, its performance must be compared to other measures of the same phenomena (remembering that the outcome of this comparative validation process will depend on which dimensions and criteria are selected). Comparison measures can be selected on the basis of three considerations.

First, there may be a benchmark measure, one that has been used frequently across assessment occasions and samples and is considered a standard against which any new measure should be compared. Examples might include measures from the BDI (see review in Katz, Katz, & Shaw, 1999), the Child Behavior Checklist (see review in Achenbach, 1996), the Hare Psychopathy Scale (see review in Rogers, 2001), the MMPI (see review in Graham, 2000), the Marital Satisfaction Inventory (see review in Snyder & Costin, 1994), or the Marital Interaction Coding System (Floyd, O'Farrell, & Goldberg, 1987). A measure that is considered to be a benchmark may still have significant limitations, and it is those limitations that can lead to the development of a new measure (e.g., the MMPI and BDI were considered flawed benchmarks, leading to the development of the MMPI–II and BDI–II).

Second, lacking a gold standard or benchmark measures, a new measure can be compared to several commonly used measures. For example, a new measure of marital satisfaction and adjustment could be compared to several commonly used measures, such as the Locke–Wallace, the Dyadic Adjustment Scale, and the Marital Satisfaction Questionnaire (see reviews in Floyd et al., 1987).

Third, there may be a simple or easily acquired measure of the same phenomena. For example, a 25-item clinician rating measure of aggression risk for psychiatric inpatients could be compared to a simple measure of patients' recent history of aggression. A 25-item self-report measure to predict treatment adherence could be compared to a single item asking clients "How likely are you to carry out treatment recommendations?" Use of a simple comparison measure is congruent with Dawes's (1999) recommendation that "zero" predictive validity should

not be the benchmark against which a new measure is compared. The incremental validity of Rorschach scores, compared with more easily acquired measures of the same phenomena, was a frequent topic in articles published in two special sections in *Psychological Assessment* (1999, 2001).

Data Analytic Strategies: Basic Concepts

The careful construction of a new measure is followed by an examination of the internal consistency, item performance characteristics (e.g., through item response theory methods), interrater and interobserver agreement, temporal stability, factor structure, content validity, and refinement and re-evaluation, as outlined in all psychometrics textbooks. This process presumably results in a homogeneous, internally consistent, and robust measure that taps the intended construct with minimal error (see Smith et al., 2003, for recommendations regarding instrument development). The next step in the incremental validation of a new measure is the application of the new measure, along with comparison and criterion measures, to an appropriate sample, to derive indices of relative and incremental predictive efficacy for the new measure. Several analytic strategies are available for these goals, depending on the dimensions of interest outlined in Table 1. For example, methodological and analytic strategies for evaluating incremental diagnostic efficacy will differ from those for evaluating treatment utility or outcome.

Whatever the incremental validity dimension of interest, there are three primary goals of these analyses: (a) to estimate the relative proportions of variance in the criterion variable(s) that can be associated with variance in the new and comparison measures, independently and in combination; (b) to estimate the proportion of variance in the criterion variables associated with variance in the new measure, above and beyond that associated with the comparison measures; and (c) to examine interaction effects associated with incremental predictive efficacy—the degree to which incremental validity of a new measure is affected by moderator variables such as sex, age, ethnicity, and diagnostic status.[3] Zero-order correlational and hierarchical regression analyses are the most useful analytic strategies to address these goals (Blais et al., 2001; Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001; Cohen & Cohen, 1983; Dawes, 1999; Strom et al., 1987; Viglione & Hilsenroth, 2001; Weiner, 2001)

Correlation Matrix

The first step in the analytic sequence is to examine a zero-order correlation matrix that includes all predictor and criterion variables

---

[3] Additional analyses would estimate the cost–benefit ratio of the new measure to inform decisions about its clinical utility; see Yates and Taub (2003), who suggested that *effectiveness* is the extent to which assessment improves the effectiveness of human services and *benefit* is the degree to which these outcomes are monetary or monetizable. If the sum of monetizable benefits of assessment exceeds the sum of the costs of treatment, the assessment can be said to be cost-beneficial.

(e.g., Dawes, 1999).[4] This matrix informs about the strength of relations among predictor variables and about the relative strength of relations between each predictor and each criterion variable. The best calculation methods (e.g.., phi, Pearson) will depend on the scaling of the variables.

Two important inferences can be derived from this zero-order correlation matrix. First, it shows the degree of collinearity, or shared variance, among the predictor variables. If the predictor variables have at least a moderate strength of relation with the criterion variables, the degree of collinearity among predictor variables suggests the degree of overlap and amount of independent information in each. The degree of overlap between two predictor variables, in turn, shows the likely increment in predictive efficacy that would occur if the two variables were combined versus if each variable were used as an independent predictor. With a high degree of collinearity (e.g., variables X_1 and X_2, correlated .9, in Table 2), the measures are redundant, and each is unlikely to show significant increases in the proportion of variance accounted for in the criterion variable when added to a regression formula that includes the other.

Inferences about the relative incremental validity of two highly correlated predictor variables are also unlikely to be robust across samples (Blais et al., 2001; Dawes, 2001). As Dawes (1999) noted, optimal weights are determined for each study and are unlikely to replicate across studies. When there is a high degree of overlap in the predictor variables, chance variations in data sets can affect the outcome of unforced and forced stepwise regression analyses with the two predictors and the criterion variable.

The zero-order correlation matrix also shows the independent strength of relations between each predictor and each criterion variable (e.g., see Row 1 in Table 2). Data on the relative validity of each predictor variable are especially useful when one best predictor of the criterion must be selected (e.g., X_3 in Table 2).

Note from Table 2 that zero-order correlations between predictor and criterion variables are insufficient to identify the best set of predictors for a criterion. The degree to which the combination of X_1 and X_2, for example, increases predictive efficacy above that associated with X_3 cannot be ascertained from this table. This incremental predictive efficacy can be estimated only through hierarchical linear regression.[5]

As Blais et al. (2001) observed, the best set of measures to predict a criterion variable are those that are highly correlated with it but not highly intercorrelated. In practical terms, zero-order correlations and hierarchical linear regression, which we discuss below in more detail, can suggest whether an assessor can use one predictor measure or if the assessor must use multiple predictors for an assessment goal.

Table 2

Sample Zero-Order Correlation Matrix of Predictor and Criterion Variables

Variable	Criterion variable (Y_1)	Predictor variables			Discriminant variable (Y_2)
		X_1	X_2	X_3	
Y_1	—	.6	.6	.8	.1
X_1			.9	.4	.3
X_2				.4	0
X_3				—	0

In validation studies with multiple criteria, the zero-order correlation matrix can also help determine the strength of relations among criteria measures. When criteria demonstrate low to moderate levels of covariation, incremental validity for a new measure must be conducted for each criterion, because generalizability of indices across criteria cannot be assumed. Of course, for most clinical assessment applications one criterion will be of primary interest, and incremental validity indices using this criterion would be weighted most heavily in decisions about the utility of a measure.

The zero-order correlation matrix can also show the independent strength of relations between each predictor and discriminant variables (Y_2).[6] Note that in Table 2, X_1 and X_2 are equally related to the criterion variable, but X_2 is unrelated to the discriminant variable. These data would be of interest when the goal for developing a new instrument is to omit items that are tapping irrelevant constructs (e.g., a new depression measure that excludes anxiety items).

Hierarchical Linear Regression

After an examination of the zero-order correlations among measures, incremental validity of a new measure is most often examined through a hierarchical linear regression analyses. These analyses address the following question: To what degree does the addition of a measure to one or more other measures increase predictive efficacy with one or more criteria? This question is answered most directly through forced stepwise regression analyses. The first step is to regress comparison measures onto the criterion. The second step is to regress the comparison plus the new measure onto the criterion. The difference in R^2 is the index of incremental validity associated with the new variable when added to the comparison measure.

In the equations that follow, $X_1 \ldots X_3$ represent gold standard, commonly used, or comparison measures; X_{10} represents the newly developed measure; and $Y_1 \ldots Y_3$ represent criterion variables. We are interested in the difference between the two R^2s (using only one comparison variable and one criterion variable in this example).

$$bX_1 = Y_1 \qquad R_1^2 \quad \text{(Step 1)} \qquad (1)$$

$$bX_1 + bX_{10} = Y_1 \quad R_2^2 \quad \text{(Step 2)} \qquad (2)$$

$$R_2^2 - R_1^2 \qquad R_{\text{diff}}^2 \qquad (3)$$

The difference between R_2^2 and R_2^1 (R_{diff}^2) is the incremental validity of R_2^2 (the new measure) compared to R_2^1 (the comparison measure) in predicting Y_1. The R_{diff}^2 difference represents the

---

[4] A *predictor variable,* as used in this article, is one used to explain variance in another variable. Temporal precedence of the predictor to the criterion variable is not necessary.

[5] It would appear that these questions could also be addressed with multiple, partial, and semipartial correlational analyses (see Clevenger et al., 2001, for an example). However, see Sechrest (1963) for a discussion of problems with this approach.

[6] A *discriminant variable,* in contrast to a *convergent variable,* is one that should not be related to the new measure. For example, measures of intelligence should not reflect (covary with) cultural or economic variables.

unique proportion of variance in Y_1 accounted for by X_{10}, above and beyond that accounted for by X_1. This difference can be subjected to an F test for statistical significance (see Dawes, 1999, for a walk-through of these analyses, using Rorschach and MMPI scales). As commonly recommended (e.g., Blais et al., 2001; Cohen & Cohen, 1983), adjusted R^2s in Steps 1 and 2 may provide a more robust estimate because they adjust the obtained R^2 to reflect sample size and the number of predictor variables.

The order of entry for predictor variables in a forced stepwise regression can be varied as a function of the goals of the analyses. If cost–benefit is a consideration, then the measure that is easiest to obtain can be entered first. The resulting R^2_{diff} can inform about the effect size benefits of using the more costly measure. The cost–benefit ratio for the three formulae can be derived by dividing the obtained R^2s by the measure of cost (see Yates & Taub, 2003). Reversing Steps 1 and 2 (comparing full and restricted multiple regression models), above, will provide identical results, because both sequences examine the independent contribution of R210 to the prediction of Y_1.

As we mentioned earlier, the appropriate analytic methods depend on whether the criterion measures are categorical or continuous. For example, Power, Doherty, et al. (1998) used logistic regression to examine the relative efficacy of two informant reports for diagnosing attention-deficit/hyperactivity disorder.

Equations 1–3, considering both orders of X variable entry, are also useful when the question being addressed is whether to add a new measure to an existing one; for example, does a combination of two measures significantly strengthen one's ability to predict aggression on a psychiatric unit? The R^2_{diff} in Equation 3 can be compared to the squared zero-order correlations (changing them to R^2 to determine the incremental predictive efficacy of two measures combined over each used independently).

Incremental predictive efficacy of a new measure, compared with an existing measure, should not be the sole basis for decisions about which measure to use. There are two reasons for caution. First, incremental criterion validity analyses, as outlined in Equations 1–3, do not indicate which of the two predictor variables is most strongly related to the criterion. This information is provided by the zero-order correlation matrix. Note that with the order of entry reversed, with the new measure first followed by the comparison measure (X_{10} followed by X_1), again presuming that both are at least moderately related to the criterion and are not highly collinear, will also result in an increment in predictive efficacy, even if the new measure is more strongly related than the comparison measure to the criterion measure. However, the degree of incremental predictive efficacy will be less than when the strongest predictor is entered last. In cases where one measure must be selected from several that are available, we are interested in the relative predictive efficacy of the measures; the one with the strongest relation to the criterion measure will often be the logical choice.

Second, a measure with incremental predictive efficacy or criterion validity can have decreased discriminant validity. Imagine a new self-report measure of depression that is more strongly related than comparison measures to a gold standard structured interview measure of depression. The incremental predictive efficacy could be due to new items that cover facets of depression that were not tapped by the comparison measures (Smith et al., 2003). However, it may also have additional items that were not included in the

comparison measures that are irrelevant to the depression construct and, consequently, may be more strongly related to the anxiety construct (i.e., have decreased discriminant validity). Equations 1–3 are applicable to discriminant validity analyses, with Y being a discriminant measure and a positive outcome being no increase in R^2 or a decreased zero-order correlation between X_{10} and discriminant measure.

The relative convergent/criterion validity and discriminant validity of a new measure depend on the degree to which the new items increase or decrease error variance in the new measure. The inclusion of many new relevant and few new irrelevant items would likely increase convergent and decrease discriminant validity, if comparison measures contained more irrelevant items. Convergent validity of the new measure would also be attenuated to the degree that the new measure contained additional inappropriate items.

Differential predictive efficacy or differential incremental validity, as a function of moderator variables, can be examined through separate stepwise regression analyses or by examining interaction effects. For example, if one wants to know the degree to which the relation between X_{10} (the new measure) and Y_1 (the criterion measure) varies as a function of comorbidity (X_5), separate correlations and R^2_{diff} can be calculated for each level of the variable (e.g., separately for males and females, for older and younger persons). The moderator can also be used as an interaction term, such that

$$bX_1 + bX_{10} + bX_5 = Y_1 \qquad \qquad R^2_3 \qquad (4)$$

$$(bX_1 + bX_{10} + bX_5) + (bX_5 \times bX_{10}) = Y_1 \quad R^2_4 \qquad (5)$$

$$R^2_3 - R^2_4 \qquad \qquad R^2_{\text{diffModer}} \quad (6)$$

where $R^2_{\text{diffModer}}$ is the incremental proportion of variance associated with interaction effects between the new measure and measures of the moderator variable, above and beyond variance associated with main effects. These data can suggest the degree to which incremental validity indices are generalizable across facets of the variable (e.g., across age groups, ethnicities, etc.). For example, Barthlow et al. (1999) found differences between men and women in the relative validity of content scales compared to clinical scales from the MMPI.

The examination of moderator effects can be a useful strategy, even when initial hierarchical regression analyses fail to support the incremental contribution of a new measure. Consider a new measure with relative predictive efficacy that is high for women but low for men. Without examining interaction effects, incremental validity indices for men and women together would obscure these differential effects.

Although less useful for most clinical judgment situations, incremental validity can also be examined for multiple criteria in combination. Equation 2 would become

$$bX_1 + bX_{10} = Y_1 + Y_2 + Y_3. \qquad (7)$$

To examine incremental validity of the new measure with a linear combination of criteria measures, the R^2 of Equation 7 would be compared to the R^2 of the equation that omits bX_{10}. The utility of this strategy for clinical judgment is limited, because the obtained R^2s reflect the relation between the best-fitting linear combination of predictors and criterion variables—the maximum correlation

with weights that are unlikely to be robust across samples and assessment occasions. It provides little information to the assessor about what measures or combination of measures should be used for a given assessment goal. Utility is also limited because clinical judgments are usually more useful when focused on specific, precisely defined constructs (Haynes & O'Brien, 2000).

Summary and Discussion: Interpretation of Incremental Validity Indices

In spite of several decades of attention from assessment scholars, the incremental validity of new clinical assessment measures is infrequently evaluated; however, incremental validation provides information that is useful to clinicians and researchers in evaluating the contribution of a new measure and when selecting one from an array of potential measures. Incremental validation can also suggest if one measure, or more than one measure used in combination, is the best strategy for a particular assessment occasion.

Inferences from incremental-validation studies are conditional. Indices of incremental validity can vary across dimensions of evaluation, such as sensitivity to change, content validity, convergent or discriminant validity, or treatment utility. Indices of incremental validity can also vary across criterion and comparison measures and across populations. Finally, decisions about whether a new and incrementally valid measure should be used are affected by cost-effectiveness considerations.

The need for a new, and presumably incrementally valid, measure is suggested by unsatisfactory performance of extant measures. Unsatisfactory performance of extant measures of a construct, across several studies, on any evaluative dimension (e.g., predictive or convergent validity, cost) or variable performance across dimensions of individual differences (e.g., age, ethnicity), suggests the need for a new measure. Insufficient content validity of measures from extant instruments is a particularly strong rationale for the development of a new measure.

We have suggested several methods of instrument development for increasing the likelihood that a new measure will be incrementally valid. The judgments that will be based on the new measure, the populations with which the measure will be used, the constructs targeted by the new measure, and alternate measures, should be clearly specified before well-articulated methods of instrument development are followed.

We have described several data analytic strategies for examining the relative and incremental validation of a new measure, with an emphasis on hierarchical and stepwise linear regression. These strategies aim to estimate the relative proportions of variance in criterion variables that can be accounted for by the new and comparison measures, independently and in combination, and to identify variables that affect those variance estimates.

One constructs a new measure and evaluates its relative and incremental validity, following the conceptual and methodological guidelines delineated above, to increase one's ability to predict important phenomena. When the process is completed, one should be able to more accurately diagnose patients, more accurately identify functional relations and construct more valid and useful case formulations, measure changes in symptoms across time more sensitively, capture multifaceted phenomena more completely, and predict important phenomena more accurately.

The index of incremental validity (e.g., R^2_{diff} of Equation 3) is necessary for judging whether the new measure should be selected for an assessment purpose. However, the R^2_{diff} is not sufficient to make that judgment; it provides only an estimate of the effect size (the relative strength of relations between the new measure and a criterion) and, along with an F test, of the robustness of the incremental validity estimate.

The decision to select or not select a new measure is also affected by three additional judgments: (a) the relative cost of acquiring data from the new measure, (b) the importance of the phenomenon predicted, and (c) the clinical utility of the new data. The smallest R^2_{diff} that is sufficient for the selection of a new measure varies directly with cost of the new measure and inversely with the importance of the criterion and the clinical utility of the data.

We reiterate these caveats to emphasize the conditional nature of inferences drawn from incremental-validation studies. Consider the incremental treatment validity of a new clinic-based, experimental functional analysis of self-injurious and stereotypic behaviors of children with autism spectrum disorders. This new assessment instrument involves systematic introduction and removal of hypothesized controlling factors (e.g., response contingencies, antecedent and setting factors) while trained observers record behavior occurrences in a time-sample format (e.g., every 15 s). A thorough functional analysis with a child with multiple factors affecting the target behaviors could require many sessions and involve multiple professionals and clinic visits, careful structure of assessment sessions, observer training, the acquisition of time-interval observation data, some distress for the child and caretakers, and a delay in beginning of standardized treatments that have been moderately effective (see reviews by Koegel, Valdez-Menchaca, Koegel, & Harrower, 2001; McClannahan, MacDuff, & Krantz, 2002; Paniagua, 2001). Incremental validity of the new experimental functional analysis in this case would primarily reflect incremental treatment utility—the degree to which data from the new experimental functional analysis improved treatment outcome compared with less costly assessment methods, such as interviewing parents and teachers about variables that affect self-injurious behavior and short-term functional analyses (e.g., one 30-min session), or compared with simply assigning the child to a standardized treatment protocol without pretreatment assessment.

Given, for example, an R^2_{diff} of .3 (a mild to moderate effect size on treatment outcome data) when the new experimental functional analysis is compared with alternative strategies (shorter, different, or no assessment), should the assessor use the new experimental functional analysis with a child with self-injurious behaviors? Alternatively, should the child be placed immediately in a standardized treatment that has been shown to be moderately effective, or should treatment be based on data from the comparison measures? Those judgments, as we have noted, will be based on several considerations. First, and most important, is the degree to which treatment outcome is improved by using the new experimental functional analysis, compared to alternative assessment strategies. The second consideration focuses on other measures of benefits. The new experimental functional analysis may be worth the added cost if it leads to a reduction in the rate of severe behavior problems, such as severe self-injury, but it may be less warranted if it leads only to a reduction of mild self-stimulatory behaviors. The third consideration is the issue of applicability

across people. For example, the new experimental functional analysis may be more warranted (exhibit better cost–benefit ratios) with children for whom standardized treatments have failed.

This is one example of how decisions about whether to use a new measure are affected by considerations in addition to incremental validity. There are many other exemplars. For example, Dawes (1999) noted that a new measure might be preferred, even if indices of incremental validity were low, if it tapped important theoretical aspects of a construct (i.e., had better content validity). Here, a small R^2_{diff} would represent an important theoretical advance. Other examples of where an important criterion might suggest that a new measure is valuable, given a small R^2_{diff}, include the prediction of suicide and recidivism in violent offenders, the prediction of aggression by patients on a psychiatric unit, and the assessment of risk of harm or neglect of children by caretakers. In another example, Strom et al. (1987) suggested that extensive (3–5 hr) neuropsychological evaluation with children with learning disabilities may be warranted only when traditional testing has resulted in inconsistent results. Glutting, Youngstrom, Ward, Ward, and Hale (1997) suggested that although factor scores on the Wechsler Intelligence Scale for Children—III (e.g., compared to full-scale IQ) increased predictive efficacy for academic achievement, the level of the increment rendered them clinically useless (i.e., the small R^2_{diff} was not offset by increased clinical utility).

These examples also illustrate the importance of establishing the incremental validity of new and existing measures, especially in assessment areas where multiple measures already exist. The unbridled proliferation of measures can lead to confusion among researchers and clinicians about which measure is best for a particular assessment purpose and the application of inappropriate instruments. With incremental-validity data, potential users can combine this information with cost, utility, and availability considerations to select the best measures. Although evaluating incremental validity may require additional time and effort by test developers, this effort is likely to be rewarded by the more consistent and informed use of clinical measures.

Although drawing the ultimate inferences from incremental validation research (should a new instrument be used) is complicated by the conditional nature of such inferences, incremental validation of new clinical assessment measures is essential for the advancement of methods and theory of clinical science, for strengthening clinical judgments, and for improving services delivered to clients.

References

Achenbach, T. M. (1996). The Child Behavior Checklist (CBCL) and related instruments. In L. I. Sederer & B. Dickey (Eds.), *Outcomes assessment in clinical practice* (pp. 97–99). Baltimore: Williams & Wilkins.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Baer, R. A., & Miller, J. (2002). Underreporting of psychopathology on the MMPI–2: A meta-analytic review. *Psychological Assessment, 14,* 16–26.

Barnett, D. W., Lentz, F. E., & Macmann, G. (2000). Psychometric qualities of professional practice. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Theory, research, and clinical foundations* (2nd ed., pp. 355–386). New York: Guilford Press.

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.

Blais, M. A., Hilsenroth, M. J., Castlebury, F., Fowler, J. C., & Baity, M. R. (2001). Predicting *DSM–IV* Cluster B personality disorder criteria from MMPI–2 and Rorschach data: A test of incremental validity. *Journal of Personality Assessment, 76,* 150–168.

Bowden, S. C., Carstairs, J. R., & Shores, E. A. (1999). Confirmatory factor analysis of combined Wechsler Adult Intelligence Scale—Revised and Wechsler Memory Scale—Revised scores in a healthy community sample. *Psychological Assessment, 11,* 339–344.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86,* 410–417.

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

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.

Dozois, D. J. A., & Dobson, K. S. (2002). Depression. In M. M. Antony & D. H. Barlow (Eds.), *Handbook of assessment and treatment planning for psychological disorders* (pp. 259–299). New York: Guilford Press.

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

Embretson, S. E., & Prenovost, L. K. (1999). Focus chapter: Item response theory in assessment research. In P. C. Kendall, J. N. Butcher, & G. N. Holmbeck (Eds.), *Handbook of research methods in clinical psychology* (2nd ed., pp. 276–294). New York: Wiley.

Floyd, F. J., O'Farrell, T. J., & Goldberg, M. (1987). Comparison of marital observational measures: The Marital Interaction Coding System and the Communication Skills Test. *Journal of Consulting and Clinical Psychology, 55,* 220–237.

Foster, S. L., & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment, 7,* 248–260.

Garb, H. N. (1985). The incremental validity of information used in personality assessment. *Clinical Psychology Review, 4,* 641–655.

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

Garb, H. N., Wood, J. M., Nezworski, M. T., Grove, W. M., & Stejskal, W. J. (2001). Toward a resolution of the Rorschach controversy. *Psychological Assessment, 13,* 433–448.

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

Graham, J. R. (2000). *MMPI–2: Assessing personality and psychopathology* (3rd ed.). New York: Oxford University Press.

Grant, A. (1938). The comparative validity of the Metropolitan readiness tests and the Pintner-Cunningham primary mental test. *Elementary School Journal, 38,* 599–605.

Hambleton, R. K., & Zaal, J. N. (Eds.). (1991). *Advances in educational and psychological testing.* Norwell, MA: Kluwer Academic.

Hargreaves, W. A., Shumway, M., Hu, T., & Cuffel, B. (1998). *Cost–outcome methods for mental health.* San Diego, CA: Academic Press.

Haynes, S. N., Nelson, K., & Blaine, D. D. (1999). Psychometric issues in assessment research. In P. C. Kendall, J. N. Butcher, & G. N. Holmbeck (Eds.), *Handbook of research methods in clinical psychology* (2nd ed., pp. 125–154). New York: Wiley.

Haynes, S. N., & O'Brien, W. B. (2000). *Behavioral assessment: A functional approach to psychological assessment.* New York: Kluwer.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity

in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238–247.

Heinrichs, R. W. (1990). Current and emergent applications of neuropsychological assessment: Problems of validity and utility. *Professional Psychology: Research and Practice, 21,* 171–176.

Heyman, R. E. (2001). Observation of couple conflicts: Clinical assessment applications, stubborn truths, and shaky foundations. *Psychological Assessment, 13,* 5–35.

Heyman, R. E., & Slep, A. M. (2003). Analogue behavioral observation. In M. Herson (Series Ed.) and S. N. Haynes & E. H. Heiby (Vol. Eds.), *Comprehensive handbook of behavioral assessment: Vol. 3. Behavioral assessment* (pp. 180–192). New York: Wiley.

Katz, J. N., Larson, M. G., Phillips, C. B., & Fossel, A. H. (1992). Comparative measurement sensitivity of short and longer health status instruments. *Medical Care, 30,* 917–925.

Katz, R., Katz, J., & Shaw, B. F. (1999). Beck Depression Inventory and Hopelessness Scale. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (2nd ed., pp. 921–933). Mahwah, NJ: Erlbaum.

Koegel, L. K., Valdez-Menchaca, M., Koegel, R. L., & Harrower, J. K. (2001). Autism. In M. Hersen & V. B. Van Hasselt (Eds.), *Advanced abnormal psychology* (2nd ed., pp. 165–190). New York: Kluwer Academic/Plenum.

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 and Biofeedback, 25,* 67–78.

McClannahan, L. E., MacDuff, G. S., & Krantz, P. J. (2002). Behavior analysis and intervention for adults with autism. *Behavior Modification, 26,* 9–26.

Messick, S. (1994). Foundations of validity: Meaning and consequences in psychological assessment. *European Journal of Psychological Assessment, 10,* 1–9.

Meyer, G. J. (Ed.). (1999). I. The utility of the Rorschach in clinical assessment [Special section]. *Psychological Assessment, 11*(3).

Meyer, G. J. (Ed.). (2001). II. The utility of the Rorschach in clinical assessment [Special section]. *Psychological Assessment, 13*(4).

Murphy-Berman, V. (1994). A conceptual framework for thinking about risk assessment and case management in child protective service. *Child Abuse and Neglect, 18,* 193–201.

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

Paniagua, F. A. (2001). Functional analysis and behavioral assessment of children and adolescents. In H. B. Vance & A. J. Pumariega (Eds.), *Clinical assessment of child and adolescent behavior* (pp. 32–85). New York: Wiley.

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

Perry, W., & Viglione, D. J. (1991). The Ego Impairment Index as a predictor of outcomes in melancholic depressed patients treated with tricyclic antidepressants. *Journal of Personality Assessment, 56,* 487–501.

Power, T. J., Andrews, T. J., Eiraldi, R. B., Doherty, B. J., Ikeda, M. J., DuPaul, G. J., & Landau, S. (1998). Evaluating attention deficit hyperactivity disorder using multiple informants: The incremental utility of combining teacher with parent reports. *Psychological Assessment, 10,* 250–260.

Power, T. J., Doherty, B. J., Panichelli-Mindel, S. M., Karustis, J. L., Eiraldi, R. B., Anastopoulos, A. D., & DuPaul, G. J. (1998). The predictive validity of parent and teacher reports of ADHD symptoms. *Journal of Psychpathology and Behavioral Assessment, 20,* 57–81.

Rogers, R. (2001). *Handbook of diagnostic and structured interviewing.* New York: Guilford Press.

Sangren, P. V. (1929). Comparative validity of primary intelligence tests. *Journal of Applied Psychology, 13,* 394–412.

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.

Silva, F. (1993). *Psychometric foundations and behavioral assessment.* Newbury Park, CA: Sage.

Smith, G. T., Fischer, S., & Fister, S. M. (2003). Incremental validity principles in test construction. *Psychological Assessment, 15,* 467–477.

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment, 7,* 300–308.

Snyder, D. K., & Costin, S. E. (1994). Marital satisfaction inventory. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 322–351). Hillsdale, NJ: Erlbaum.

Strom, D. A., Gray, J. W., Dean, R. S., & Fischer, W. E. (1987). The incremental validity of the Halstead–Reitan Neuropsychological Battery in predicting achievement for learning-disabled children. *Journal of Psychoeducational Assessment, 2,* 157–165.

Tupes, E. C. (1950). An evaluation of personality-trait ratings obtained by unstructured assessment interviews. *Psychological Monographs: General and Applied, 64,* 1–23.

Viglione, D. J., & Hilsenroth, D. J. (2001). The Rorschach: Facts, fictions, and future. *Psychological Assessment, 13,* 452–471.

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.

Weiner, I. B. (2001). Advancing the science of psychological assessment: The Rorschach inkblot method as exemplar. *Psychological Assessment, 13,* 423–432.

Wiggins, N., & Kohen, E. S. (1971). Man versus model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology, 19,* 100–106.

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.