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

provement 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, 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

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 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 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 $r_{A} = [r_{YA} -$ -c-297T* (r)Tj /F5 1 Tfu61804 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 associmanner 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 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 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 encoun-

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: 107. Construct validity and comparison with the Psychopathic Deviate scale. *Psychological Assessment, 8,* 281–2983.ofw T7622lD4i-318.5(of)-318.93.