JOURNAL OF PERSONALITY ASSESSMENT, *77*(2), 307–332

Toward More Clinically Relevant Assessment Research

Robert E. McGrath
School of Psychology
Fairleigh Dickinson University

The purpose of this article is to encourage the design and conduct of more clinically relevant assessment research. The discussion focuses on the distinction between research design elements that are appropriate to the examination of validity versus utility issues in assessment. This was addressed in 2 ways. The first part of the article summarizes the results of an informal methodological review of 108 empirically based articles published in 3 major assessment journals (*Assessment, Journal of Personality Assessment,* & *Psychological Assessment*) during the period of September 1998 through August 1999. The results indicated that studies published in these journals demonstrated appropriate sensitivity to some aspects of utility but not others. For example, none of the articles addressed the process of interpretation in clinical settings or the reactions of stakeholders to assessment. Recommendations for enhancing the integration of validity and utility issues in the design of research are discussed. These include methodological issues and topics in need of further study. A particularly important example of the latter is the need for research identifying factors that foster positive impressions of psychological assessment.

Over the last 5 years there has been a dramatic increase in the number of articles recognizing the need for more clinically relevant psychotherapy research (e.g., Clarke, 1995; Garfield, 1996; Goldfried & Wolfe, 1996; Hoagwood, Hibbs, Brent, & Jensen, 1995; Hollon, 1996; Persons & Silberschatz, 1998). In the context of psychotherapy outcomes, this discussion has often revolved around the distinction between efficacy and effectiveness research. Efficacy research is conducted to demonstrate that therapeutic gains occur specifically because of participation in the treatment of interest. Internal validity is maximized through rigorous control of potential confounds. In contrast, effectiveness research attempts to estimate the outcomes for therapeutic methods as they are naturally practiced. In effectiveness

TABLE 1
Characteristics of Efficacy Versus Effectiveness Research

Efficacy Research	*Effectiveness Research*
Random assignment to treatments	Patient selects some treatment parameters (e.g., provider, modality)
Strict experimental control over threats to internal validity	Threats to internal validity controlled statistically if at all
Outcome variables are specific and reliably measured	Outcome measures are global and attitudinal (e.g., global ratings of satisfaction)
Manual-based treatment	Individually tailored intervention
Comparison to placebo	No placebo group
Equalization of groups on length of treatment	Variable treatment length
Raters blinded to treatment	Data generated by participants in treatment (therapists and patients)
Strict exclusionary criteria (e.g., dually diagnosed patients excluded)	Inclusionary

studies, experimental control is sacrificed for ecological validity. Efficacy research is used to evaluate whether psychotherapy can work; effectiveness research is used to evaluate whether it does work (Mook, 1983).

Table 1 summarizes some of the methodological distinctions between the two types of research (see Seligman, 1995). Although studies typically combine elements of both in their design, for many years psychotherapy researchers were encouraged by funding agencies to emphasize strict control to rule out alternative explanations for therapeutic change. However, clinicians complained that the results gave them little guidance about whether treatments worked in the real world (Persons & Silberschatz, 1998).

In response to these objections, the psychotherapy research community has worked to change its ways, culminating in a recent decision by the National Institute of Mental Health to modify its funding guidelines to foster more research examining the effectiveness of treatment (Foxhall, 2000; Norquist, Lebowitz, & Hyman, 1999). Demonstrating the efficacy of a treatment is no longer considered sufficient grounds for justifying its use.

The distinction between conceptual research investigating whether a clinical activity can work, and practical research investigating how it functions in the real world, is one that assessment researchers were aware of long before the current interest among therapy researchers. In the context of assessment research, though, this distinction is usually couched in terms of the validity versus utility of an instrument (Hayes, Nelson, & Jarrett, 1987; Wiggins, 1973).

There are strong parallels between psychotherapy research on efficacy and assessment research on validity; the same could be said for effectiveness research

and utility research. The goal of validity research is to evaluate the extent to which an assessment method measures what it is supposed to measure. Validity is traditionally treated as a formal and inherent characteristic of the instrument that determines its functioning across situations. Although validity is a necessary condition for ensuring that the use of an instrument could be worthwhile, it does not ensure its use will be worthwhile. Utility research has to do with the practical value of the instrument in applied settings; it evaluates the value added by the use of the instrument in a particular context. Validity has to do with whether an instrument can work, utility with whether it does work.

VALIDITY AND UTILITY RESEARCH

Table 2 contrasts methodological elements that characterize the study of validity versus utility. This is not intended as an exhaustive list, but it does capture some of the major differences. The following sections explain each of these differences.

Population

The essential question for a validity study is whether an instrument measures the psychological state or characteristic it is supposed to measure. This question can have testable implications for both clinical and nonclinical populations. However, only research with the former provides direct evidence about the value of the instrument in clinical settings. For example, demonstrating that college students' scores

TABLE 2
Characteristics of Validity-Focused Versus Utility-Focused Research

Validity Research	*Utility Research*
Independent of population	Specific to populations served in applied settings
Validity is maximized by quantitative predictors and criteria	Utility is best estimated by categorical predictors and criteria
Test results are evaluated independent of clinician interpretation	Clinician interpretation of the test is important
Zero-order relationships are most important	Incremental validity is important
Validity is maximized by balanced base rates	Utility is best estimated by naturally occurring base rates
Preferred conditional probabilities are sensitivity and specificity	Preferred conditional probabilities are positive and negative predictive power
Interviews are structured to improve reliability	Interviews are generally unstructured
Raters are trained to improve reliability	Ratings are based on clinical experience
Perceptions of the testing are not relevant	Perceptions of the testing are important
Costs and benefits of testing are irrelevant	Cost–benefit ratios are central issues

on a measure of anxiety increase during midterms corroborates the validity of the instrument. However, the results do not offer guidance to the clinician about how useful the instrument will be for predicting the severity of anxiety symptoms in a clinical population.

Quantitative Versus Categorical Variables

Because most assessment instruments generate quantitative scores, the most straightforward way to evaluate validity is by correlating those scores with data generated using some alternate method of tapping the construct. This practice, however, is uncharacteristic of the clinical setting in several ways.

First, it ignores the central role of decision making in practical assessment. Assessments are usually conducted for purposes of classification: Is this person suicidal or not? A qualified job applicant or not? Direct evaluation of an instrument's capacity to produce valid decisions requires first dichotomizing cases based on test results, then correlating predicted status with actual status. This approach places additional burdens on the researcher. Not only does it require the additional step of defining a method of classification based on test scores, it is also more likely to result in statistical tests that are not significant (Cohen, 1983).

Clinical Interpretation

Second, the focus on bivariate relationships in validity research also ignores the normal interpretive process. In practical settings, assessment instruments are almost always interpreted by the clinician in the context of potential moderators. At the least, this includes biographical or contextual information, but can also include the behavior of the test taker during the testing and performance on other measures. The combination of multiple data sources makes unusual or even unique combinations of outcomes possible. At this point, the clinician falls back on the "art" of assessment, trying to make sense of what can seem to be incongruent or even contradictory information in the absence of nomothetic guidelines. Ideally, if the instrument is likely to be interpreted in light of other data sources, clinically relevant research would attempt to reflect that practice.

Incremental Validity Tests

Third, the use of bivariate relationships to evaluate validity ignores the existence of alternate scales measuring the same construct. This is appropriate: A test's validity is not impacted by the existence of other methods for measuring the same construct. In practice, though, the clinician usually has a choice between several measures. More clinically useful research evaluates the relative or incremental validity of an instrument as it compares to alternatives. Despite reference

to it as a form of validity, incremental validity is more closely tied with issues of utility than validity. It is specific to the alternate data sources that are likely to be available, and the assessment questions for which the instrument is likely to be considered appropriate.

Base Rates

When predictors or criteria are categorical, the issue of base rates becomes relevant. In some studies, the base rate for one or more of the variables is set so that the number of participants who are positive and the number who are negative are approximately equal. Before the spread of computers, this was often done because formulas for hand computation of statistics are simpler when frequencies are balanced. Today, the equalization of base rates seems unjustified except in some experimental designs. For example, analog research on the effectiveness of response style indicators usually compares the test scores of participants who complete the measure under standard instructions to a group of participants who complete the measure under instructions to respond in an invalid manner. In a review of studies using the Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1943) to identify faking bad (Rogers, Sewell, & Salekin, 1994), many of the studies compared groups of approximately equal size, particularly those using repeated measures designs. In contrast, Rogers, Sewell, and Goldstein (1994) estimated the actual base rate for malingering to be 15.7% in forensic and 7.4% in nonforensic settings. It is worth noting that asymmetrical base rates can dramatically reduce the power of significance tests (Dawes, 1993).

Preferred Conditional Probabilities

The inclusion of dichotomous predictors and criteria allow for the computation of two sets of conditional probabilities: sensitivity and specificity, and positive predictive power (PPP) and negative predictive power (NPP). Sensitivity and specificity represent the probability of a positive or negative outcome on the test given to the test taker's population. Sensitivity is the probability that the test will correctly identify test takers who are positive for the target condition. Specificity is the probability the test will correctly identify test takers who are negative for the condition.

PPP and NPP represent the probability of the test taker's population given the test outcome. PPP is the probability that test takers are positive if the test outcome is positive. NPP is the probability that test takers are negative if the test outcome is negative.

There does not seem to be any absolute advantage to either set of conditional probabilities for the purpose of evaluating test validity. However, sensitivity and specificity have traditionally been preferred in studying the formal aspects of tests.

Two factors likely contribute to this bias. First, sensitivity and specificity should be less sensitive to variations in base rates across samples. Second, PPP and NPP require a Bayesian perspective on probability, the philosophical implications of which have traditionally been considered problematic by some statisticians (Oakes, 1986). Specifically, how can population membership be treated as conditional on test outcome when it is presumed to be an unconditional characteristic of the individual?

Despite technical concerns about the statistics, PPP and NPP are more consistent with the prospective nature of clinical decision making. Testing usually occurs precisely because the test taker's status in unknown, so that the only variable available to the assessor is the test outcome. The PPP and NPP provide clinically useful information about the degree of confidence the assessor may attach to a clinical decision based on that outcome.

Interview and Rater Standardization

A common strategy for enhancing the control of methodological construct validity is the use of standardized interviewing and rating methods. Standardization increases the potential for reliable ratings, reducing the potential for alternate explanations of findings (particularly in instances where tests are not significant), and increasing effect sizes. This last advantage is problematic in the context of utility research, however. Because interviews and ratings made in the real world are usually not standardized, standardization for research purposes can potentially lead to overestimates of the effect sizes to be expected in clinical practice.

Perceptions of Testing

There are several classes of stakeholders associated with psychological testing, test administrators, test takers, recipients of test results, and third-party payers being the most important. Stakeholders' perceptions of testing are considered an inadequate barometer of validity, with good reason. Several lines of research raise concerns about whether even clinicians can judge the accuracy of test data. These include research on the acceptance of Barnum statements as interpretive feedback (Dickson & Kelly, 1985; but see Schroeder & Lesyk, 1976), on common sources of error in judgment (Garb, 1998), and on the perception of illusory correlations between test data and criteria (Chapman & Chapman, 1969).

Appropriate suspicions about stakeholder perceptions as a measure of validity may have led us to neglect the role those perceptions play in determining the usefulness of an assessment. Put most simply, assessment is useless if the results are not used in subsequent decision making. Experience suggests that no matter how valid the results are, individuals who receive the feedback must find the results interesting or useful before they are likely to use them.

Cost–Benefit Analysis

The cost of using an instrument is immaterial to its validity. In contrast, the balance between the perceived costs (financial and practical) associated with an instrument's administration, scoring, and interpretation, and the perceived benefits associated with the outcomes may be the single most important determinant of any technique's clinical use. The empirical analysis of cost–benefit ratios can be a particularly daunting endeavor. At times, some costs or benefits can be perceived rather than actual, as in the case of illusory correlations. At other times, costs or benefits may be difficult to identify or quantify. For example, one could hypothetically find that clinicians choose a familiar instrument over a newer, potentially more valid instrument because learning the new instrument represents a significant cost factor.

FOCUSING ON UTILITY

As with efficacy and effectiveness, assessment studies often combine elements of both validity and utility research, and validity is a prerequisite for utility. Even so, discussions with clinicians involved in the practice of assessment indicate they perceive researchers as overly focused on conceptual validation issues, at the expense of practical usefulness. Researchers for their part tend to find it difficult to design ecologically valid studies with sufficient methodological rigor to provide clear results.

This article is intended as a call for greater consideration of clinical relevance in the design of assessment research. The remainder of the article focuses on two issues. First, recent studies in major assessment journals were reviewed for how well they incorporated utility concerns into the design. Second, based on the results of the review, several suggestions are made for improving the focus on utility issues in future assessment research.

REVIEW OF THE LITERATURE

To demonstrate the point that utility issues are often neglected in assessment research, I conducted an informal review of selected assessment journals. This review was not intended to be exhaustive or to capture the current state of assessment research. It was conducted to demonstrate ways in which the clinical value of research could potentially be improved. Accordingly, no attempt was made to ensure the reliability of the analysis. The Appendix provides a complete summary of the results should one wish to examine them.

I reviewed 1 year's worth of issues from three journals that are considered important publication outlets for the assessment research community. These included *Assessment* (four issues), *Journal of Personality Assessment* (six issues), and *Psy-*

chological Assessment (four issues). The issues selected covered the period of September 1998 through August 1999, the most recent at that time.

Across the 14 issues, there were 137 articles, excluding book reviews. Of these, 6 were reanalyses of previously presented data such as meta-analyses, and another 23 were conceptual articles. This left 108 articles that provided individual results from one study or a series of studies.

I then reviewed each of these articles for methodological choices that reflect the extent to which the authors attempted to duplicate the clinical setting. The results of this review are provided in Table 3. The table begins with those dimensions that showed a reasonable mix between validity and utility concerns.

TABLE 3
Presence of Methodological Features That Enhance Clinical Relevance

Feature	*No.*	*%*
Sample Constitution		
Nonclinical		
Nonclinical scale	15	13.9
Clinical scale	17	15.7
Norming–surveying	4	3.7
With general clinical	11	10.2
With specific clinical	9	8.3
Clinical only	52	48.2
Predictors[a]		
Categorical	6	9.1
Mixed	22	33.3
Quantitative	38	57.6
Criteria[b]		
Categorical	30	45.5
Mixed	14	21.2
Quantitative	22	33.3
Base rates[b]		
Natural	26	39.4
Manipulated	17	25.8
Unclear	3	4.5
No categorical predictors or criteria	20	30.3
Predictive power[b]		
Provided	15	22.7
Could be computed	11	16.7
Could not be computed	0	0.0
No categorical predictors or no categorical criteria	40	60.6
Interview[c]		
Yes	18	69.2
Unstructured	7	26.9
Unclear	1	3.9

(continued)

TABLE 3 (Continued)

Feature	*No.*	*%*
Ratings[d]		
Raters trained for study or reviewed	22	32.8
Raters untrained and not reviewed	27	40.3
Unclear	18	26.9
Perceptions of stakeholders		
Evaluated	1	0.9
Not evaluated	107	99.1
Clinical interpretation evaluated		
Yes	1	0.9
No	107	99.1
Test validity comparison[b]		
Direct test	14	21.2
Nonstatistical comparison	14	21.2
No comparison	38	57.6
Cost–benefit analysis of testing		
Yes	0	0.0
No	108	100.0

[a]Among prediction studies; $N = 66$. [b]Among prediction studies. [c]Among studies with interviews; $N = 26$. [d]Among studies with ratings; $N = 67$.

Sample Constitution

The first factor considered was the composition of the sample. Many of the articles summarized multiple studies, and the nature of the sample often varied across studies. For example, some articles described both the development of an instrument and its initial validation. Others included both a structural analysis of the instrument (e.g., a factor analysis) and an analysis of its relation with other variables. All of these studies had implications for the validity of the instrument. However, only those studies that examined the effectiveness of the instrument as a predictor of an external criterion (which will be referred to as *prediction studies* here) provided a direct parallel to the manner in which the test is likely to be used clinically. Therefore, if the article included studies that examined the relationship between the target instrument and other variables, classification was based on those portions of the article only. If the article included no prediction component, classification was based on all the studies.

Six types of sample were identified:

1. Normals (nonclinical individuals) completed a test that is appropriate to the population (e.g., a personality measure).
2. Normals completed an instrument only likely to be used with pathological populations (e.g., a neuropsychological measure or a measure of psychopathology).

3. Normals participated in a normative sample or a survey (and there was no prediction component).
4. Normals were compared to a general clinical sample.
5. Nonclinical individuals were compared to a specific clinical sample (usually having a specific diagnosis).
6. Clinical patients only.

All but the third condition can be used to evaluate the validity of an instrument directly. However, only the first and last conditions can be used to estimate the effect sizes likely to be found in common applied settings.

The results demonstrate diversity in the types of samples used. Approximately one half were limited to clinical patients, and when studies that included any clinical patients were included, the percentage increased to 66.7%. However, the administration of a clinical instrument to nonclinical individuals is probably the least useful condition from the perspective of evaluating clinical utility, and yet was the second most common sample scenario.

Quantitative or Categorical Variables

The next five dimensions apply only to the 66 articles that included a prediction study. The first issue addressed was whether the predictors (scores on the target instrument) were dichotomized indicators of respondent status. In more than one half of the cases, predictors were represented only in a quantitative form. In one third of cases, there was a mixture of quantitative and categorical predictors. Frequently, this mix occurred because the same predictor was used in both its original quantitative and dichotomized forms for different analyses. In only 9% of studies were categorical predictors used exclusively.

In contrast, almost one half of the studies included only categorical criteria. This was typical because the target variable was naturally dichotomous, such as the presence or absence of a diagnosis or clinical state. The use of quantitative rather than categorized predictors in such instances is particularly questionable, given that the corresponding clinical task is classification. Taking findings for the predictors and criteria together, the tendency seems to be to analyze variables in their original format. In 20 out of 66 prediction studies, all the predictors and criteria were quantitative. In 40 out of 66, either the predictors or the criteria were all quantitative.

Base Rate Manipulation

The next variable examined was whether the base rates of categorical predictors or criteria were manipulated in some way. Out of 66 prediction studies, the most common decision was to use naturally occurring base rates. However, 17 studies used

manipulated base rates that equated group membership, and 20 had no categorical predictors or criteria.

Predictive Power

If both the predictor and criterion are dichotomous, the PPP and NPP can be computed. The 26 studies meeting this condition were reviewed to determine whether the authors provided the PPP and NPP of the instrument, or at least provided enough information so they could be computed. In the most consistent finding indicating an awareness of utility issues, every one of the studies did so. However, tempering the positive nature of this finding is the large number of studies without dichotomized variables.

The results bear comparison to those published previously by Kessel and Zimmerman (1993), who reviewed the reporting of conditional probabilities in 26 studies published between 1980 and 1991 on the prediction of depression with self-report measures. They found only 7 out of 27 studies reported predictive power statistics (25.9%), compared to 15 out of 26 studies in the this sample (57.7%). Given that one of the two journals they reviewed, *Psychological Assessment,* was also used in this study, it seems reasonable to hypothesize that the difference is largely due to increased awareness among researchers of the importance of predictive power.

Standardization

Standardized interviewing procedures were used in the majority of studies that involved interviews. However, this statistic was skewed by those studies specifically conducted to evaluate the standardized interview. When clinicians were engaged in typical clinical activities, as when the purpose of the interview was to judge whether the test taker met inclusionary criteria, a mix of standardized and unstandardized interviews was used. In contrast, less than one half of patient ratings were completed by trained raters, or raters whose conclusions were reviewed for accuracy. In many circumstances, the ratings represented actual field data (e.g., based on chart reviews).

Perceptions of Stakeholders

The final four findings were the most disappointing in terms of establishing utility. Only one study evaluated the perceptions of stakeholders about testing, and that study had to do with clinicians' attitudes toward computer-based interpretive systems. There were no studies published in these journals evaluating how test takers or recipients of test results felt about the assessment.

Clinical Interpretation

Similarly, only one study investigated how clinicians interpret tests in practice. This was a study of clinicians' perception of the "pull" from Thematic Apperception Test cards. There were no empirical studies in any of the major assessment journals during this period investigating the process by which clinicians integrate information from multiple data sources into a coherent set of judgments about the test taker.

Incremental Validity

Among the prediction studies, only 28 evaluated whether the target instrument had a statistical advantage over alternative methods, about 42% of the 66 studies. In those studies that compared multiple measures of the same criterion, one half did not conduct a statistical test of their relative effectiveness. Usually, the authors provided bivariate correlations or other statistics for both predictors and commented on the differences. Only 14 studies directly evaluated incremental validity. This was most commonly accomplished using hierarchical regression, but also included a handful of studies that tested for differences between correlations or used other techniques. It is interesting to note that not one study evaluated whether the target instrument was superior to asking the test taker for a simple rating of intensity or status (see Burisch, 1984).

Cost–Benefit Analysis

Finally, not one study in the three primary assessment journals attempted to define the conditions under which the benefits of using an assessment method could justify its costs. This was true even though some of the instruments investigated are quite time consuming, such as some of the interview-based, neuropsychological, or projective measures.

DISCUSSION

In some cases, the results demonstrate a balance between validity and utility concerns. A good example of this occurs with the base rate issue. The preference for natural base rates suggests sensitivity to ecological validity. However, the frequency of manipulated base rates reflects the continued importance of using experimental methods at times to generate more clearly interpretable results.

In general though, the results suggest little importance is placed on utility issues. In some cases this may occur for the sake of enhancing validity, as when trained raters are used. In other cases though, decisions about design probably had

more to do with expediency or practicality than anything else, as when normals were administered measures of pathology.

Some of the more neglected issues in the list are also ones that may be the most important for the future of psychological testing. Surveys have consistently indicated the use of psychological tests is declining. This decline can be directly linked to the routine rejection of reimbursement requests for assessment (Eisman et al., 1998; Griffith, 1997). Convincing managed care organizations and other gatekeepers of the value of assessment has become a key factor in assuring its future viability.

Two aspects of test utility are particularly crucial to this effort. First, stakeholders must personally perceive assessment as valuable. Second, direct cost savings available through the use of psychological assessment must be demonstrated. If the personal reservations of stakeholders can be alleviated, and if it can be demonstrated that assessment results in a cost savings, gatekeepers will find it difficult to defend present policy.

Enhancing Personal Perceptions

Recent work on therapeutic assessment offers some clues about methods for improving stakeholders' perceptions of testing. By allowing the test taker an active role in determining the purposes of the assessment and interpreting the results, therapeutic assessment impacts positively on test takers' impressions of and benefits from the testing (Finn & Tonsager, 1992; Newman & Greenway, 1997). It has also been found effective as a means of reducing the likelihood of early termination and enhancing the therapeutic alliance (Ackerman, Hilsenroth, Baity, & Blagys, 2000). This in turn can enhance utilization reviewers' impressions of the process (Finn & Martin, 1997).

These findings suggest several important research questions relevant to enhancing the perceptions of stakeholders in general. First, it suggests that the process of the assessment may be as important as, if not more important than, the validity of the interpretation for enhancing perceptions of the assessment. Second, this principle may be useful for directly managing the perceptions of other stakeholders besides the test taker.

Despite the lack of research on this second topic, personal experience suggests it may be so. Assessment often occurs in collaborative settings, such as inpatient or forensic units, where psychologists work with psychiatrists, probation officers, and other caregivers. It is frequently the unfortunate case in such settings that the assessment is not integral to service provision. At the extreme, the psychologist is called to the unit to conduct the assessment, assesses the individual without further input from unit staff, and presents the results as a report when the testing is completed. The problem with this model is that, without input from the test taker or

feedback recipient, and without an explanation of how the results can be interpreted in context, the findings are often of little value to unit staff. Other professionals are left without a sense of how psychological testing can contribute directly to their goals.

The alternative is a collaborative model that parallels the therapeutic assessment model. The assessment begins with a dialogue between the assessor and persons requesting the assessment to determine goals. In addition to generating a report focusing on those goals, the psychologist collaborates with staff on the interpretation of results and their implications for treatment, and encourages further requests for information. It is a participatory model, in which the results of the assessment are used as part of the ongoing process of determining treatment and evaluating psychosocial issues. If the results are valid, other professionals can be left with a positive sense of the potential value for assessment. This is potentially reflected in future willingness to request testing, attitudes expressed to students, and ultimately on the viability of assessment in that setting.

Cost–Benefit Analysis

Attitudes toward assessment play an important role in determining policy decisions about psychological testing, but the demonstrated balance between the costs and benefits of testing are equally important. For example, the consistent finding that decisions based on the mechanical combination of test data is superior to clinical judgment (Grove, Zald, Lebow, Snitz, & Nelson, 2000) seems to have had little impact on policy in the areas of disability evaluation and managed mental health (Meyer et al., 1998). The reason is simple: An advantage in validity is unimportant to policymakers in these situations if there is no demonstrated advantage in outcomes or cost savings. Given the importance of cost–benefit demonstrations to convincing policymakers about the value of assessment, it is surprising how little research has appeared on this topic in recent years (for a review, see Meyer et al., 1998). In fact, the most recent significant contribution to this literature seems to be Hayes et al.'s (1987) discussion of the treatment utility of assessment. In contrast, the corresponding concept in psychotherapy, usually referred to medical cost offset, continues to receive a great deal of attention (Chiles, Lambert, & Hatch, 1999).

CONCLUSIONS

Like efficacy and effectiveness, demonstrating validity and utility represent complementary research goals. Angoff (1988) concluded that references to the validity or the reliability of an instrument without reference to context are overly abstract to the point of being unrealistic. Validity represents the necessary precondition for utility, whereas validity without the consideration of utility is empty.

The review of the literature tends to support clinicians' contention that researchers neglect issues of practical relevance. This neglect may reflect the greater emphasis on issues of validity in the training of researchers, rather than practical barriers to the design of more ecologically valid research. The following represent several ways in which researchers could easily incorporate more naturalistic elements into their research:

1. Research studies that include both nonclinical and clinical samples can be enhanced by providing effect size estimates based both on comparisons between the two samples and between subgroups within the clinical sample.
2. Variables should be analyzed in both quantitative and dichotomous forms. The former would be more relevant to evaluating the validity of the original scale, whereas the latter would provide more reasonable effect size estimates for clinical users. Requiring analysis of dichotomized variables would also force test developers to consider appropriate cut scores for all new quantitative measures.
3. Base rates are sometimes manipulated for methodological reasons, as in some experimental studies. Even so, random subsamples can be generated to estimate effect sizes at a more realistic base rate level as well. Requiring researchers to do so would force them to address what would be reasonable effectiveness of the instrument in settings where it is likely to be used.
4. If there are multiple efficient measures available for the same construct, any investigation of validity should be expected to include the analysis of incremental validity.
5. When a new instrument or procedure is intended for clinical use, scale developers should be expected to identify clinical situations where adding the new instrument or using it to replace existing measures may reasonably be found to offer a cost–benefit advantage over existing methods. It is important to note that a cost–benefit advantage need not require the demonstration of incremental validity. For example, in settings where clinical decisions sometimes have dramatic consequences, as in some forensic situations, simply corroborating the findings of another instrument may justify the cost.
6. Sometimes the natural clinical situation is so complicated it cannot be replicated given the limited resources available to researchers, or the results would be so messy that they are uninterpretable. Researchers can have a bad habit of equating messy research with unpublishable research. This should not be the case, as long as the researcher explicitly addresses the complexity of the situation, justifies the compromises chosen between internal and external validity, and demonstrates the importance of studying the topic. Complex situations suggest the need for multiple studies, with varying degrees of emphasis on validity and utility issues, before final conclusions are possible.
7. Although the importance of effect sizes estimates and confidence intervals in addition to significance test results are generally recognized, they are particu-

larly important for establishing the clinical value of an instrument. Thinking in terms of classification problems, clinicians need to know that an instrument can increase their confidence in a conclusion. They also need to know how much of an increase it potentially offers. The size and variability of the effect is essential for making judgments about how well an instrument performs clinically.

In addition to expanding our research methods and analyses, though, it is also important to expand the domain of appropriate research goals. It may be that the future of assessment depends to a large degree on how people who are not assessors perceive assessment. The development of testable models for involving professional stakeholders in the assessment experience, and for demonstrating the cost–benefit advantages of testing, represents a valuable target for future research. In a world in which clinicians often perceive researchers as out of touch with the reality of assessment and mental health decisions are increasingly driven by the perceived value of an intervention, such studies may prove the greatest gift possible for researchers to give to clinicians.

REFERENCES

Ackerman, S. J., Hilsenroth, M. J., Baity, M. R., & Blagys, M. D. (2000). Interaction of therapeutic process and alliance during psychological assessment. *Journal of Personality Assessment, 75,* 82–109.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 19–32). Hillsdale NJ: Lawrence Erlbaum Associates, Inc.

Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist, 39,* 214–227.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74,* 271–280.

Chiles, J. A., Lambert, M. J., & Hatch, A. L. (1999). The impact of psychological interventions on medical cost offset: A meta-analytic review. *Clinical Psychology: Science and Practice, 6,* 204–220.

Clarke, G. N. (1995). Improving the transition from basic efficacy research to effectiveness studies: Methodological issues and procedures. *Journal of Consulting and Clinical Psychology, 63,* 718–725.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249–253.

Dawes, R. M. (1993). Prediction of the future versus an understanding of the past: A basic asymmetry. *American Journal of Psychology, 106,* 1–24.

Dickson, D. H., & Kelly, I. W. (1985). The "Barnum effect" in personality assessment: A review of the literature. *Psychological Reports, 57,* 367–382.

Eisman, E. J., Dies, R. R., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Meyer, G. J., & Moreland, K. L. (1998). *Problems and limitations in the use of psychological assessment in contemporary healthcare delivery: Report of the Board of Professional Affairs Psychological Assessment Work Group, Part 2.* Washington, DC: American Psychological Association.

Finn, S. E., & Martin, H. (1997). Therapeutic assessment with the MMPI–2 in managed health care. In J. N. Butcher (Ed.), *Personality assessment in managed care* (pp. 131–152). New York: Oxford University Press.

Finn, S. E., & Tonsager, M. E. (1992). Therapeutic effects of providing MMPI–2 test feedback to college students awaiting therapy. *Psychological Assessment, 4,* 278–287.

Foxhall, K. (2000, July/August). Research for the real world: NIMH is pumping big money into effectiveness research to move promising treatments into practice. *Monitor on Psychology, 31,* 28–36.

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

Garfield, S. L. (1996). Some problems associated with "validated" forms of psychotherapy. *Clinical Psychology: Science and Practice, 3,* 218–229.

Goldfried, M. R., & Wolfe, B. E. (1996). Psychotherapy practice and research: Repairing a strained alliance. *American Psychologist, 51,* 1007–1016.

Griffith, L. F. (1997). Surviving no-frills mental healthcare: The future of psychological assessment. *Journal of Practical Psychiatry and Behavioral Health, 3,* 255–258.

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.

Hoagwood, K., Hibbs, E., Brent, D., & Jensen, P. (1995). Introduction to the special section: Efficacy and effectiveness in studies of child and adolescent psychotherapy. *Journal of Consulting and Clinical Psychology, 63,* 683–687.

Hollon, S. D. (1996). The efficacy and effectiveness of psychotherapy relative to medications. *American Psychologist, 51,* 1025–1030.

Kessel, J. B., & Zimmerman, M. (1993). Reporting errors in studies of the diagnostic performance of self-administered questionnaires: Extent of the problem, recommendations for standardized presentation of results, and implications for the peer review process. *Psychological Assessment, 5,* 395–399.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Moreland, K. L., Eisman, E. J., & Dies, R. R. (1998). *Benefits and costs of psychological assessment in healthcare delivery: Report of the Board of Professional Affairs Psychological Assessment Work Group, Part 1*. Washington, DC: American Psychological Association

Mook, D. G. (1983). In defense of external validity. *American Psychologist, 38,* 379–387.

Newman, M. L., & Greenway, P. (1997). Therapeutic effects of providing MMPI–2 test feedback to clients at a university counseling service: A collaborative approach. *Psychological Assessment, 9,* 122–131.

Norquist, G., Lebowitz, B., & Hyman, S. (1999). Expanding the frontier of treatment research. *Prevention and Treatment, 2, Article 0001a.* Retrieved August 13, 2001 from the World Wide Web: journals.apa.org/prevention/volume2/pre0020001a.html

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences.* Chichester, England: Wiley.

Persons, J. B., & Silberschatz, G. (1998). Are results of randomized controlled trials useful to psychotherapists? *Journal of Consulting and Clinical Psychology, 66,* 126–135.

Rogers, R., Sewell, K. W., & Goldstein, A. (1994). Explanatory models of malingering: A prototypical analysis. *Law and Human Behavior, 18,* 543–552.

Rogers, R., Sewell, K. W., & Salekin, R. T. (1994). A meta-analysis of malingering on the MMPI–2. *Assessment, 1,* 227–237.

Schroeder, H. E., & Lesyk, C. K. (1976). Judging personality assessments: Putting the Barnum report in perspective. *Journal of Personality Assessment, 40,* 470–474.

Seligman, M. E. P. (1995). The effectiveness of psychotherapy: The *Consumer Reports* study. *American Psychologist, 50,* 965–974.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment.* Reading, MA: Addision-Wesley.

APPENDIX

Article[a]	*Sample*	*Predictors Categorical*	*Criteria Categorical*	*Base Rate*	*PPP*	*Perceptions of Testing Evaluated*	*Interview Structured or Semi-Structured*	*Raters Trained or Reviewed*	*Evaluated Clinician Interpretation*	*Validity Compared*	*Cost–Benefit Ratio*
Jourrnal of Personality Assessment, 73(1)											
Strong	Patients only	No predictors	No criteria			No			No		No
Cramer	Normals, nonpathology test	None	Some	Manipu-lated		No		Yes	No	No	No
Cumella	Patients only	Some	All	Natural	Could be computed	No	No	No	No	No	No
Leichsenring	Normal and general patient groups	Some	All	Manipu-lated	Could be computed	No	Yes	Yes	No	No	No
Hiatt	Patients only	Some	Some	Natural	Provided	No	No	No	No	No	No
Bing	Normals, nonpathology test	None	None			No			No	Statistically	No
Ornduff	Patients only	None	All	Natural		No	No	No	No	No	No
Lynam	Normals, pathology test	None	None			No			No	No	No
Podar	Normal and specific patient groups	None	All	Manipu-lated		No			No	No	No
Journal of Personality Assessment, 72(2)											
Strassle	Normals, norming or surveying	No predictors	No criteria			No		No	Yes		No
Meyer	Normal and general patient groups	No predictors	No criteria			No		Unclear	No		No
Carbone	Patients only	None	All	Natural		No	No	No	No	Statistically	No
Martin	Normals, nonpathology test	No predictors	No criteria			No			No		No

Rouse	Normals, pathology test	No predictors	No criteria			No		No		No
Baer	Normal and general patient groups	Some	All	Manipu-lated	Provided	No		No	Non-statistically	No
Journal of Personality Assessment, 72(1)										
Meyer	Patients only	No predictors	No criteria			No	Unclear	No		No
Osberg	Patients only	None	None			No		No	Non-statistically	No
Vieth	Normals, nonpathology test	No predictors	No criteria			No		No		No
Collins	Normal and specific patient groups	No predictors	No criteria			No		No		No
Cramer	Normal and general patient groups	None	All	Manipu-lated		No	Unclear	No	No	No
Baity	Patients only	None	Some	Natural		No	Yes	No	Statistically	No
Romm	Patients only	No predictors	No criteria			No		No		No
Journal of Personality Assessment, 71(3)										
Schinka	Normal and general patient groups	No predictors	No criteria			No		No		No
Holaday	Patients only	No predictors	No criteria			No	No	No		No
Liljequist	Normal and specific patient groups	Some	All	Manipu-lated	Could be computed	No	No	No	No	No
Guevara	Normals, nonpathology test	None	None			No		No	No	No
Savard	Patients only	None	Some	Natural		No		No	No	No

(continued)

APPENDIX (Continued)

Article[a]	*Sample*	*Predictors Categorical*	*Criteria Categorical*	*Base Rate*	*PPP*	*Perceptions of Testing Evaluated*	*Interview Structured or Semi-Structured*	*Raters Trained or Reviewed*	*Evaluated Clinician Interpretation*	*Validity Compared*	*Cost–Benefit Ratio*
Litinsky	Patients only	None	All	Natural		No		Yes	No	No	No
Edens	Normals, pathology test	Some	All	Manipulated	Provided	No			No	No	No
McNulty	Patients only	All	None	Natural		No		No	No	Nonstatistically	No
Porcerelli	Normals, pathology test	None	All	Manipulated		No		Yes	No	No	No
Silberg	Patients only	None	All	Manipulated		No		No	No	No	No
Journal of Personality Assessment, 71(2)											
Winter	Normals, nonpathology test	Some	Some	Unclear	Could be computed	No		Unclear	No	Nonstatistically	No
Ritzler	Patients only	No predictors	No criteria			No		No	No		No
Clemence	Normal and general patient groups	Some	All	Manipulated	Provided	No		No	No	Statistically	No
Rapport	Normals, norming or surveying	No predictors	No criteria			No			No		No
King	Normals, pathology test	Some	Some	Natural	Could be computed	No			No	Statistically	No
Rolland	Normals, nonpathology test	No predictors	No criteria			No			No		No
Personality Assessment, 11(2)											
Osman	Patients only	Some	All	Natural	Provided	No			No	Statistically	No
Weathers	Patients only	All	Some	Unclear	Provided	No	Yes	Unclear	No	Nonstatistically	No
Goldberg	Normals, pathology test	None	None			No			No	No	No

Van Gerwen	Patients only	None	None			No	No	No	No	No	No
Hardy	Normals, pathology test	Some	Some	Natural	Provided	No	Yes	Yes	No	No	No
Ruehlman	Normals, pathology test	None	None			No	Yes	Unclear	No	No	No
Megargee	Patients only	No predictors	No criteria			No			No		No
Caruso	Normals, nonpathology test	No predictors	No criteria			No		Yes	No		No
Ryan	Normals, nonpathology test	No predictors	No criteria			No		Yes	No		No
Personality Assessment, 11(1)											
Cooke	Normal and general patient groups	No predictors	No criteria			No	Unclear	Yes	No		No
Handwerk	Patients only	No predictors	No criteria			No		No	No		No
Bagby	Patients only	All	None	Natural		No			No	Nonstatis-tically	No
Barthlow	Patients only	None	None			No		No	No	Statistically	No
Chambless	Patients only	No predictors	No criteria			No	Yes	Yes	No		No
Carroll	Patients only	None	None			No	Yes	Yes	No	Nonstatis-tically	No
Cacciola	Patients only	No predictors	No criteria			No	Yes	Yes	No		No
Stein	Patients only	Some	All	Natural	Provided	No		No	No	Statistically	No
Rouse	Patients only	Some	All	Natural	Provided	No		No	No	Statistically	No
Hayes	Patients only	None	None			No		Yes	No	No	No
Personality Assessment, 10(4)											
Butler	Patients only	None	None			No	Yes	Yes	No	Nonstatis-tically	No

(continued)

APPENDIX (Continued)

Article[a]	*Sample*	*Predictors Categorical*	*Criteria Categorical*	*Base Rate*	*PPP*	*Perceptions of Testing Evaluated*	*Interview Structured or Semi-Structured*	*Raters Trained or Reviewed*	*Evaluated Clinician Interpretation*	*Validity Compared*	*Cost–Benefit Ratio*
Cocco	Patients only	Some	Some	Natural	Could be computed	No	Yes	Unclear	No	Nonstatistically	No
Orbach	Normal and general patient groups	None	Some	Natural		No			No	No	No
Poythress	Patients only	Some	Some	Natural	Provided	No	Yes	No	No	No	No
Campbell	Patients only	None	None			No		No	No	Nonstatistically	No
Otto	Patients only	None	Some			No	Yes	No	No	No	No
Personality Assessment, 10(3)											
Endler	Normal and specific patient groups	None	All	Natural		No			No	No	No
Foa	Normal and specific patient groups	None	Some	Natural		No	Yes	Yes	No	No	No
Bryant	Patients only	All	All	Natural	Could be computed	No	No	No	No	No	No
Arbisi	Patients only	Some	All	Manipulated	Provided	No			No	Statistically	No
Trull	Normal and general patient groups	None	None			No	Yes	Yes	No	Statistically	No
Cole	Normals, nonpathology test	No predictors	No criteria			No			No		No
Power	Normals, pathology test	Some	All	Natural	Provided	No	Yes	Yes	No	Statistically	No
Cole	Normals, pathology test	No predictors	No criteria			No			No		No

Vanderploeg	Normal and specific patient groups	Some	All	Manipu-lated	Could be computed	No	Yes	No	Nonstatis-tically	No
Canivez	Normals, nonpathology test	No predictors	No criteria			No	No	No		No
Ward	Patients only	No predictors	No criteria			No		No		No
Assessment, 6(2)										
Kurtz	Normals, nonpathology test	No predictors	No criteria			No		No		No
Harlan	Normals, pathology test	None	None			No		No	No	No
Gladsjo	Normals, norming or surveying	No predictors	No criteria			No	Unclear			No
Wiegner	Patients only	No predictors	No criteria			No	Unclear	No		No
Lyons	Normals, nonpathology test	None	None			No		No	No	No
Assessment, 6(1)										
Zinn	Normals, pathology test	None	None			No	Yes	No	No	No
Anderson	Normals, pathology test	None	None			No		No	No	No
Campbell	Normals, pathology test	No predictors	No criteria			No		No		No
Guilmette	Patients only	None	None			No	Unclear	No	Nonstatis-tically	No
Allen	Normals, pathology test	No predictors	No criteria			No	Unclear	No		No
Archer	Patients only	No predictors	No criteria			No		No		No
Putzke	Patients only	No predictors	No criteria			No	Yes	No		No

(continued)

APPENDIX (Continued)

Article[a]	*Sample*	*Predictors Categorical*	*Criteria Categorical*	*Base Rate*	*PPP*	*Perceptions of Testing Evaluated*	*Interview Structured or Semi-Structured*	*Raters Trained or Reviewed*	*Evaluated Clinician Interpretation*	*Validity Compared*	*Cost–Benefit Ratio*
McMinn	Normals, norming or surveying	No predictors	No criteria			Yes			No		No
Vittengl	Normal and specific patient groups	No predictors	No criteria			No	Yes	No	No		No
Livingston	Patients only	No predictors	No criteria			No		Unclear	No		No
Assessment, 5(4)											
Lubin	Normal and specific patient groups	Some	All	Manipu-lated	Could be computed	No		No	No	No	No
Janus	Patients only	No predictors	No criteria			No			No		No
Shedler	Patients only	None	None			No		No	No	No	No
Schnirman	Normals, pathology test	No predictors	No criteria			No		Unclear	No		No
Demsky	Normals, nonpathology test	None	None			No		Unclear	No	Nonstatis-tically	No
Allen	Patients only	No predictors	No criteria			No	Yes	No	No		No
Diehr	Normal and general patient groups	No predictors	No criteria			No			No		No
Wasyliw	Patients only	None	All	Natural		No			No	No	No
Rogers	Patients only	Some	All	Manipu-lated	Provided	No	Yes	Unclear	No	No	No
Maruish	Patients only	No predictors	No criteria			No			No		No
Assessment, 5(3)											
Morey	Normal and general patient groups	Some	All	Manipu-lated	Provided	No			No	Nonstatis-tically	No

Culbertson	Patients only	No predictors	No criteria			No		Unclear	No		No
Trahan	Patients only	All	All	Unclear	Could be computed	No		Unclear	No	No	No
Mills	Patients only	None	Some	Natural		No	No	No	No	No	No
Wetzler	Patients only	Some	All	Natural	Provided	No		No	No	Statistically	No
Gutentag	Normal and specific patient groups	None	All	Manipu-lated		No		Yes	No	Statistically	No
Merritt	Normals, pathology test	All	All	Natural	Could be computed	No	Yes	Yes	No	No	No
Mroczek	Normals, nonpathology test	No predictors	No criteria			No			No		No
Meyers	Patients only	None	All	Manipu-lated		No		Unclear	No	No	No

Note. *Journal of Personality Assessment, 72*(3) consisted solely of nonempirical articles. PPP = positive predictive power.
[a]First author only.

Robert McGrath
School of Psychology
Fairleigh Dickinson University
Teaneck NJ 07666
E-mail: mcgrath@alpha.fdu.edu

Received July 31, 2000
Revised January 11, 2001