Can There Be Validity Without Reliability?

PAMELA A. MOSS

Reliability has traditionally been taken for granted as a necessary but insufficient condition for validity in assessment use. My purpose in this article is to illuminate and challenge this presumption by exploring a dialectic between psychometric and hermeneutic approaches to drawing and warranting interpretations of human products or performances. Reliability, as it is typically defined and operationalized in the measurement literature (e.g., American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1985; Feldt & Brennan, 1989), privileges standardized forms of assessment. By considering hermeneutic alternatives for serving the important epistemological and ethical purposes that reliability serves, we expand the range of viable high-stakes assessment practices to include those that honor the purposes that students bring to their work and the contextualized judgments of teachers.

Educational Research, Vol. 23, No. 2, pp. 5–12.

Some time ago, I submitted to a journal a manuscript in which my coauthors and I argued for the value of teachers' contextualized judgments in making consequential decisions about individual students and educational programs. Drawing on epistemological strategies typically used by qualitative or interpretive researchers, we offered an example of how teachers' narrative evaluations of their students' collected work, which varied in substance from student to student and classroom to classroom, might be warranted and used for accountability purposes. We based our argument for the value of this sort of contextualized assessment on the unique quality of information it might provide when used in conjunction with more standardized forms of assessment and on the educational benefits it might have for teachers and students. We warranted the narrative evaluation, in part, in critical dialogue among readers about the multidimensional evidence contained in students' folders and, in part, in documentation of evidence allowing subsequent readers of the report to "audit" or confirm the conclusions for themselves.

Reviewer B thought our manuscript was a "superb and important article" and gave it her "highest endorsement." Reviewer A thought the manuscript should not be published in its current form because we had "confused the purpose of assessment with that of instruction" and had "failed to establish reliability" (in this case, adequate consistency among independent readings). She commented that our argument showed a lack of understanding of the essential function of reliability, not only in service of validity, but also "for fairness to the student to prevent the subjectivity and potential bias of an individual teacher." The editor, faced with the dilemma of divergent opinions, wrote that he would be willing to publish an article based on the manuscript if it dealt effectively with the concerns of Reviewer A. He commented, diplomatically, that he feared our position "might be misread as a rejection of a fundamental measurement principle." He noted that "any measurement should have adequate reliability for its purposes, otherwise it is not good measurement, regardless of its positive features."

There is an instructive irony embedded in this anecdote. The process by which a working decision was reached regarding our manuscript was based in an epistemology that more closely resembled the one we had proposed than the one against which our manuscript was evaluated. The editor's decision was not grounded in the consistency among independent readings, which diverged substantially; rather, he made a thoughtful judgment based upon a careful reading of both sets of comments and his own evaluation of the manuscript. I am confident that he was concerned with the validity and fairness of his decision. Of course, I didn't agree with his initial decision, but our dialogue continued through the mail, and the paper improved (and was published) as we strengthened our argument in response to his concerns. I am also confident that both the readers of the journal and I were well served by the written dialogue that accompanied what, for me, was a "high-stakes" decision.

My purpose in this article is to illuminate and challenge the presumption that reliability, as it's typically defined and operationalized in the professional measurement literature (e.g., AERA et al., 1985; Feldt & Brennan, 1989), is essential to sound assessment practice; in doing this, I give particular attention to the context of accountability in public education. I explore a dialectic between two diverse approaches to drawing and warranting interpretations of human products and performances—one based in psychometrics and one in hermeneutics. This task, I believe, honors Messick's (1989) proposed "Singerian" mode of inquiry in validity research, where one inquiring system is observed in terms of another inquiring system "to elucidate the distinctive technical and value assumptions underlying each system application and to integrate the scientific and ethical implications of the inquiry" (p. 32). My point is not to overturn a traditional criterion but rather to suggest that it be treated as only one of several possible strategies of serving important epistemological and ethical purposes. The choice among reliability and its alternatives has conse-

PAMELA A. MOSS *is assistant professor, University of Michigan, 4220 School of Education, 610 East University, Ann Arbor, MI 48109-1259. She specializes in educational measurement and evaluation.*

quences for stakeholders in the educational system. That choice should be not be taken for granted or treated as nonproblematic.

Concerns With the Traditional View of Reliability

"Without reliability, there is no validity." Many of us who develop and use educational assessments were taught to take this maxim for granted as a fundamental principle of sound measurement. *The Standards for Educational and Psychological Testing* (AERA et al., 1985), along with most major measurement texts (e.g., Crocker & Algina, 1986; Cronbach, 1990), present reliability as a necessary, albeit insufficient, condition for validity. Theoretically, reliability is defined as "the degree to which test scores are free from errors of measurement.... Measurement errors reduce the reliability (and therefore the generalizability) of the score obtained for a person from a single measurement" (AERA et al., 1985, p. 19). Typically, reliability is operationalized by examining consistency, quantitatively defined, among independent observations or sets of observations that are intended as interchangeable—consistency among independent evaluations or readings of a performance, consistency among performances in response to independent tasks, and so on. In fact, Feldt and Brennan (1989) describe the "essence" of reliability analysis as the "quantification of the consistency and inconsistency in examinee performance" (p. 105). In this article, I focus primarily on issues of reliability or generalizability across tasks (products or performances by the person or persons about whom conclusions are drawn) and across readers (interpreters or evaluators of those performances).

Less standardized forms of assessment, such as performance assessments, present serious problems for reliability, in terms of generalizability across readers and tasks as well as across other facets of measurement. These less standardized assessments typically permit students substantial latitude in interpreting, responding to, and perhaps designing tasks; they result in fewer independent responses, each of which is more complex, reflecting integration of multiple skills and knowledge; and they require expert judgment for evaluation. Empirical studies of reliability or generalizability with performance assessments are quite consistent in their conclusions that (a) reader reliability, defined as consistency of evaluation across readers on a given task, can reach acceptable levels when carefully trained readers evaluate responses to one task at a time and (b) adequate task or "score" reliability, defined as consistency in performances across tasks intended to address the same capabilities, is far more difficult to achieve (e.g., Breland, Camp, Jones, Morris, & Rock, 1987; Dunbar, Koretz, & Hoover, 1991; Shavelson, Baxter, & Gao, 1993). In the case of portfolios, where the tasks may vary substantially from student to student and where multiple tasks may be evaluated simultaneously, inter-reader reliability may drop below acceptable levels for consequential decisions about individuals or programs (Koretz, McCaffrey, Klein, Bell, & Stecher, 1992; Nystrand, Cohen, & Martinez, 1993).[1]

Validity researchers in performance assessment, building on the pioneering work of Messick (1964, 1975, 1980, 1989) and Cronbach (1980, 1988) that expanded the definition of validity to include consideration of social consequences, have stressed the importance of *balancing* concerns about reliability, replicability, or generalizability with additional criteria such as "authenticity" (Newmann, 1990), "directness" (Frederiksen & Collins, 1989), or "cognitive complexity" (Linn, Baker, & Dunbar, 1991). This balancing of often competing concerns has resulted in the sanctioning of lower levels of reliability, as long as "acceptable levels are achieved for particular purposes of assessment" (Linn et al., 1991, p. 11; see Messick, 1992, and Moss, 1992, for a review). Where acceptable levels have not been reached, recommendations for enhancing reliability without increasing the number of tasks or readers beyond cost-efficient levels have typically involved (a) increasing the specification of tasks or scoring procedures, thereby resulting in increased standardization, and (b), in the case of portfolios, disaggregating the contents so that tasks may be scored, independently, one task at a time. Wiley and Haertel (in press) offer a promising means of addressing task reliability without the constraining assumption of homogeneity of tasks. As part of a comprehensive assessment development process, they suggest carefully analyzing assessment tasks to describe the capabilities required for performance, scoring tasks separately for the relevant capabilities, and examining reliability within capability across tasks to which the capability applies. While this supports the use of complex and authentic tasks that may naturally vary in terms of the capabilities elicited, it still requires detailed specification of measurement intents, performance records, and scoring criteria. So although growing attention to the consequences of assessment use in validity research provides theoretical support for the move toward less standardized assessment, continued reliance on reliability, defined as quantification of consistency among independent observations, requires a significant level of standardization.

Given the growing body of evidence about the impact of high-stakes assessment on educational practice (Corbett & Wilson, 1991; Johnston, Weiss, & Afflerbach, 1990; Smith, 1991), this privileging of standardization is problematic. As Resnick and Resnick (1992) conclude, to the extent that assessment results "are made visible and have consequences" (p. 55), efforts to improve performance on a given assessment "seem to drive out most other educational concerns" (p. 58). There are certain intellectual activities that standardized assessments can neither document nor promote; these include encouraging students to find their own purposes for reading and writing, encouraging teachers to make informed instructional decisions consistent with the needs of individual students, and encouraging students and teachers to collaborate in developing criteria and standards to evaluate their work. A growing number of educators are calling for alternative approaches to assessment that support collaborative inquiry and foreground the development of purpose and meaning over skills and content in the intellectual work of students (Greene, 1992; Willinsky, 1990) and teachers (Darling-Hammond, 1989; Lieberman, 1992). If Resnick and Resnick (1992) are correct in their conclusion that what isn't assessed tends to disappear from the curriculum, then we need to find ways to document the validity of assessments that support a wider range of valued educational goals. And, as Wolf, Bixby, Glenn, and Gardner (1991) have suggested, we need to "revise our notions of high-agreement reliability as a cardinal symptom of a useful and viable approach to scoring student performance" and "seek other sorts of evidence that responsible judgment is unfolding" (p. 63).

Unquestionably, reliability serves an important purpose. Underlying our concerns about reliability are both epistemological and ethical issues. These include the extent to which we can generalize to the construct of interest from particular samples of behavior evaluated by particular readers and the extent to which those generalizations are fair. There are, however, alternative means of serving those purposes. The decision about which strategy to use should depend upon the aims and consequences of the assessment in question. In the sections that follow, I explore the potential of a hermeneutic approach to drawing and warranting interpretations of human products or performances.[2] Although the focus here is on reliability (consistency among independent measures intended as interchangeable), it should be clear that reliability is an aspect of construct validity (consonance among multiple lines of evidence supporting the intended interpretation over alternative interpretations). And as assessment becomes less standardized, distinctions between reliability and validity blur.

A Hermeneutic Approach to Interpretation

Hermeneutics characterizes a general approach to the interpretation of meaning reflected in any human product, expression, or action, often referred to as a text or "text analog." Although hermeneutics is not a unitary tradition, most hermeneutic philosophers share a holistic and integrative approach to interpretation of human phenomena that seeks to understand the whole in light of its parts, repeatedly testing interpretations against the available evidence until each of the parts can be accounted for in a coherent interpretation of the whole. (Edited volumes by Bleicher, 1980, and Ormiston and Schrift, 1990, provide excerpts from the works of the major philosophers in the hermeneutic tradition.) Recently, a number of philosophers of science have suggested that this "hermeneutic circle" of initial interpretation, validation, and revised interpretation characterizes much that occurs in the natural as well as the social sciences (Bernstein, 1983; Diesing, 1991; Kuhn, 1986).

Hermeneutic writings are often categorized into three major perspectives reflecting differences in the relative "authority" they give to text, context, and reader in building an interpretation (Bleicher, 1980; Ormiston & Schrift, 1990; Rabinow & Sullivan, 1987; Warnke, 1987). One perspective, reflected in the writings of Schleiermacher, Dilthey, Betti, and, more recently Hirsch, treats "hermeneutical theory" as a methodology intended to produce relatively objective or correct interpretations that reflect the original intent of the author while bracketing the preconceptions of the reader. Here, the hermeneutic circle is conceived of in terms of a dialectical relationship between the parts of the text and the whole. A second perspective, "hermeneutic philosophy," which is reflected in the writings of Heidegger and Gadamer, recognizes that the reader's preconceptions, "enabling" prejudices, or foreknowledge are inevitable and valuable in interpreting a text. In fact, they make understanding possible. Here, the hermeneutic circle is viewed as including a dialectic between reader and text to develop practically relevant knowledge. A third perspective, often called "critical" or "depth hermeneutics," reflected in the writings of Habermas and Apel, highlights the importance of considering social dynamics that may distort meaning. Here the hermeneutic circle expands to include a critique of ideology from the ideal perspective of an unconstrained communication—one in which all parties concerned (including researcher and researched) approach each other as equals. These differing perspectives provide alternative resolutions to concerns about such issues as subjectivity, objectivity, and generalizability that psychometricians have confronted in building *their* interpretations.

Comparing Hermeneutic and Psychometric Approaches

Major differences between hermeneutic and psychometric approaches to validity research can be characterized largely in terms of how each treats the relationships between the parts of an assessment (individual products or performances) and the whole (entire collection of available evidence) and how human judgment is used to arrive at a well-warranted conclusion.

In a typical psychometric approach to assessment, each performance is scored independently by readers who have no additional knowledge about the student or about the judgments of other readers. Inferences about achievement, competence, or growth are based upon composite scores, aggregated from independent observations across readers and performances, and referenced to relevant criteria or norm groups. These scores, whose interpretability and validity rest largely on previous research, are provided to users with guidelines for interpretation. Users are typically (and appropriately) advised to consider scores in light of other information about the individual, although mainstream validity theory provides little guidance about how to combine such information to reach a well-warranted conclusion—a task to which hermeneutic analysis is well suited.

A hermeneutic approach to assessment would involve holistic, integrative interpretations of collected performances that seek to understand the whole in light of its parts, that privilege readers who are most knowledgeable about the context in which the assessment occurs, and that ground those interpretations not only in the textual and contextual evidence available, but also in a rational debate among the community of interpreters. Here, the interpretation might be warranted by criteria like a reader's extensive knowledge of the learning context; multiple and varied sources of evidence; an ethic of disciplined, collaborative inquiry that encourages challenges and revisions to initial interpretations; and the transparency of the trail of evidence leading to the interpretations, which allows users to evaluate the conclusions for themselves.

Illustrations of More Hermeneutic Approaches to Assessment

In higher education, we have a number of models of high-stakes assessments that are not standardized. For instance, consider how we confer graduate degrees, grant tenure, or, as my introductory paragraphs illustrated, make decisions about which articles will be published and which will not. As an extended example, consider what, based on my experience, appears to be a typical process for making decisions about hiring faculty colleagues. Candidates submit portfolios of their work and evaluations by others. While there may be some minimal standardization (e.g., three representative publications, teaching evaluations, and a letter articulating their programs of research and teaching), candidates are expected to compile evidence that they believe best represents the substance and quality of their

work. Search committee members are selected because of their areas of expertise and affiliation—to cover the knowledge and political bases that a thoughtful decision requires. They are not trained to agree on a common set of criteria and standards; rather, it is expected that all will bring expertise to bear in evaluating candidates' credentials. Candidates' portfolios, interviews, and presentations are not parceled out to be evaluated independently by different committee members. Rather, each member examines all the evidence available to reach and support an integrative judgment about the qualifications of the candidates. These judgments are not aggregated to arrive at a set of scores; rather, they are brought to the table for a sometimes contentious discussion. Disagreement is taken seriously and positions are sometimes changed as different perspectives are brought to bear on the evidence. The final decision represents a consensus or compromise based on that discussion. An ethic of fairness typically pervades these discussions, as credentials are viewed and reviewed to make sure no qualified candidates have been overlooked. The recommendation, rationale, and supporting evidence are passed on for review to other levels of the system, typically including executive committees, administrators, and affirmative action committees. Taken together, these procedures serve to warrant the validity and fairness of the decision. Would we really believe the process more fair and valid if we followed more traditional assessment techniques of evaluating the parts independently and aggregating the scores?

These models are not unheard of in K–12 education. Edited volumes by Berlak et al. (1992) and Perrone (1991) describe successful examples of more contextualized and dialogic forms of assessment. For instance, at the Walden School in Racine, Wisconsin, students prepare papers or exhibits that are evaluated by a committee of teachers, parents, other students, and members of the community. At the Prospect School in North Bennington, Vermont, teachers meet regularly to discuss progress of individual students or curriculum issues, in much the same way that physicians conduct case conferences. In pilot projects in England, committees of teachers, supervisors, and others at the school level engage in periodic audits of the individual portfolios, and committees at higher levels of the system review the procedures of the school level committees to ensure that appropriate standards are being followed. (Elsewhere, my colleagues and I have provided an extended illustration; Moss et al., 1992.)

While the above examples focus primarily on individual assessment, other examples more directly address the problems of providing system level information in more dialogic and contextualized forms. In Pittsburgh, Pennsylvania, the Arts PROPEL project has involved committees of teachers in designing a districtwide portfolio assessment system and has invited educators from outside the district to audit the portfolio evaluation process (LeMahieu, Eresh, & Wallace, 1992). At the Brooklyn New School in New York City, the staff has developed a "learner centered accountability system" in which a comprehensive set of structures and processes have been set up to support opportunities for guided student choice, collaborative learning and inquiry among teachers and administrators, active involvement of students and their families in educational decisions, and regular involvement by educators and researchers from outside the school in formative evaluation activities (Darling-Hammond & Snyder, 1992). These examples of assessment are all consistent with what Darling-Hammond and Snyder call a professional model of accountability, which seeks evidence that teachers are engaging in collaborative inquiry to make knowledge-based decisions that respond to individual students' needs.

Yes, But What About Generalizability and Fairness?

Regardless of whether one is using a hermeneutic or psychometric approach to drawing and evaluating interpretations and decisions, the activity involves inference from observable parts to an unobservable whole that is implicit in the purpose and intent of the assessment. The question is whether those generalizations are best made by limiting human judgment to single performances, the results of which are then aggregated and compared with performance standards, or by expanding the role of human judgment to develop integrative interpretations based on all the relevant evidence.

With a psychometric approach, generalizability is warranted in quantitative measures of consistency across independent observations (across tasks, readers, and so on). As I argued above, the nature of the warrant privileges more standardized forms of assessment. When operationalized in this way, inadequate consistency puts the validity of the assessment use in jeopardy. While consistency or consensus supports the validity of the interpretations in both psychometric and hermeneutic approaches, the difference rests in how it is addressed. Here I will consider the way generalizations may be constructed and warranted from more hermeneutic perspectives and how this, in turn, expands possibilities for assessment.

Generalization Across Tasks

With respect to generalization across tasks, the goal of a more hermeneutic approach is to construct a coherent interpretation of the collected performances, continually revising initial interpretations until they account for all of the available evidence. Inconsistency in students' performance across tasks does not invalidate the assessment. Rather, it becomes an empirical puzzle to be solved by searching for a more comprehensive or elaborated interpretation that explains the inconsistency or articulates the need for additional evidence. A well-documented report describes the evidence available to other readers so that they may judge its adequacy for themselves in supporting the desired generalization. Moreover, when the interpretation informs a subsequent action, such as a revised pedagogical strategy, the success of the action becomes another warrant of the validity of the working interpretation. This is consistent with the characterization of the hermeneutic circle by Packer and Addison (1989) as a dialectic between problem and solution that furthers the concern of the reader.

In terms of task selection, hermeneutic approaches to assessment can allow students and others being assessed substantial latitude in selecting the products by which they will be represented—a latitude that need not be constrained by concerns about quantitative measures of consistency across tasks. As my hiring illustration suggested, permitting those assessed to choose products that best represent their strengths and interests may, in some circumstances, enhance not only validity but also fairness. With psychometric approaches to assessment, fairness in task selection

has typically been addressed by requiring that all subjects respond to equivalent tasks, which have been investigated for bias against various groups of concern (Cole & Moss, 1989). Neither approach ensures fairness: With the psychometric approach, we may present students with tasks for which there is differential familiarity, and with the hermeneutic approach, students may not be prepared to choose the products that best represent their capabilities. However, both approaches to fairness in task selection are defensible and deserve discussion.

Generalization Across Readers

With respect to generalization across readers, a more hermeneutic approach to assessment would warrant interpretations in a critical dialogue among readers that challenged initial interpretations while privileging interpretations from readers most knowledgeable about the context of assessment. Initial disagreement among readers would not invalidate the assessment; rather, it would provide an impetus for dialogue, debate, and enriched understanding informed by multiple perspectives as interpretations are refined and as decisions or actions are justified. And again, if well documented, it would allow users of the assessment information, including students, parents, and others affected by the results, to become part of the dialogue by evaluating (and challenging) the conclusions for themselves.

Concerns about the objectivity (and hence the fairness) of such a process have been thoughtfully addressed by qualitative researchers from both hermeneutic and postpositivist empirical traditions of research. Phillips (1990), a persuasive defender of postpositivist empirical research, citing Scriven's (1972) distinction between quantitative and qualitative senses of objectivity, notes that consensus or agreement among independent observations is no guarantor of objectivity. Rather, he defines objectivity, procedurally, as acceptance of a critical tradition: "The community of inquirers must be a critical community, where dissent and reasoned disputation (and sustained efforts to overthrow even the most favored of viewpoints) are welcomed as being central to the process of inquiry" (pp. 30–31). Moreover, he notes, objectivity is no guarantor of "truth":

> A critical community might never reach agreement over, say, two viable alternative views, but if both of these views have been subjected to critical scrutiny, then both would have to be regarded as objective.... And even if agreement is reached, it can still happen that the objective view reached within such a community will turn out to be wrong—this is the cross that all of us living in the new nonfoundationalist age have to learn to bear! (p. 31)

This dialogic perspective on the role of the critical community of interpreters in an age where no knowledge is viewed as certain is consistent with the recent writing of Cronbach (1988, 1989) and Messick (1989) on the philosophy of validity. It is also consistent with the writing of hermeneutic philosophers. Here, however, a comparison among the hermeneutic perspectives that I described earlier reflects instructive differences in the role of the readers' preconceptions and the role of the power dynamics within the social context when interpretations are formed. Proponents of hermeneutic philosophy and depth hermeneutics would question the possibility of "objective" knowledge that required readers to bracket their preconceptions. Bernstein (1983), citing Gadamer, argues that we cannot bracket all our prejudices because there is no knowledge or understanding without prejudice (foreknowledge). (Imagine trying to interpret a response written in an unknown foreign language.) The point is to discriminate between blind and enabling prejudices by critically testing them in the course of inquiry.

In a very real sense, attention to reliability actually works against critical dialogue, at least at one phase of inquiry. It leads to procedures that attempt to exclude, to the extent possible, the values and contextualized knowledge of the reader and that foreclose on dialogue among readers about the specific performances being evaluated. A hermeneutic approach to assessment encourages such dialogue at all phases of the assessment. As Bernstein (1983) argues, the absence of a sure foundation against which to test knowledge claims does not condemn us to relativism: Themes in the work of Gadamer, Habermas, and others writing in the hermeneutic and critical traditions look to "dialogue, conversation, undistorted communication, communal judgment, and the type of rational wooing that can take place when individuals confront each other as equals" (p. 223).

If interpretations are warranted through critical dialogue, then the question of who participates in the dialogue becomes an issue of power, as proponents of critical or depth hermeneutics would remind us. In articulating criteria for valid assessment in the service of accountability purposes, a number of assessment specialists have explicitly advised against using the judgments of classroom teachers (e.g., Mehrens, 1992; Resnick & Resnick, 1992). Resnick and Resnick, for instance, assert:

> A principal requirement of accountability and program evaluation tests is that they permit detached and impartial judgments of students' performance, that is, judgments by individuals other than the students' own teachers, using assessment instruments not of the teachers' devising.... Like accountability tests, selection and certification tests must be impartial. The public function of certification would not be met if teachers were to grade the performance of their own students. (pp. 48–50)

In contrast, other educators raise concerns about the absence of teachers' voices in mechanisms of accountability that affect them and their students (e.g., Darling-Hammond & Snyder, 1992; Erickson, 1986; Lieberman, 1992). Erickson, for instance, laments the fact that teachers' accounts of their own practices typically have no place in the discourse of schooling. He notes that in other professions, including medicine, law, and social work, "it is routine for practitioners to characterize their own practice, both for purposes of basic clinical research and for the evaluation of their services" (p. 157) and that "the lack of these opportunities [for teachers] is indicative of the relative powerlessness of the profession outside the walls of the classroom" (p. 157). Similar concerns have been raised about the role of students in assessments that have consequences in their lives (e.g., Greene, 1992; Willinsky, 1990).

From a psychometric perspective, the call for "detached and impartial" high-stakes assessment reflects a profound concern for fairness to individual students and protection of stakeholders' interests by providing accurate information. From a hermeneutic perspective, however, it can be criticized as arbitrarily authoritarian and counterproductive,

because it silences the voices of those who are most knowledgeable about the context and most directly affected by the results. Quantitative definitions of reliability locate the authority for determining meaning with the assessment developers. In contrast, Gadamer (cited in Bernstein, 1983) argues that the point of philosophical hermeneutics is to correct "the peculiar falsehood of modern consciousness: the idolatry of scientific method and of the anonymous authority of the sciences" (p. 40) and to vindicate "the noblest task of the citizen—decision-making according to one's own responsibility—instead of conceding that task to the expert" (p. 40).

Of course, the validity of any consequential interpretation, including the extent to which it is free from inappropriate or "disabling" prejudices, must be carefully warranted through critical, evidence-based review and dialogue. The process proposed is not dissimilar from the way decisions are made and warranted in the law (see Ricoeur, 1981). Again, neither a psychometric nor a hermeneutic approach guarantees fairness; however, a consideration of the assumptions and consequences associated with both approaches leads to a better informed choice.

Implications

I now return to my title, "Can there be validity without reliability?" When reliability is defined as consistency among *independent* measures intended as *interchangeable*, the answer is, yes. Should there be? Here, the answer is, it depends on the context and purposes for assessment. My argument shares much with Mishler's (1990) views on reliability as a means of warranting knowledge claims:

> Reformulating validation as the social discourse through which trustworthiness is established elides such familiar shibboleths as reliability, falsifiability, and objectivity. These criteria are neither trivial nor irrelevant, but they must be understood as particular ways of warranting validity claims rather than as universal, abstract guarantors of truth. They are rhetorical strategies ... that fit only one model of science. (p. 420)

Like Mishler, I am not advocating the abandonment of reliability. Rather, I am advocating that we consider it one alternative for serving important epistemological and ethical purposes—an alternative that should always be justified in critical dialogue and in confrontation with other possible means of warranting knowledge claims. As Messick (1989) has advised, such confrontations between epistemologies illuminate assumptions, consequences, and the values implied therein. Ultimately, the purpose of educational assessment is to improve teaching and learning. If reliability is put on the table for discussion, if it become an option rather than a requirement, then the possibilities for designing assessment and accountability systems that reflect a full range of valued educational goals become greatly expanded.

I believe the dialogue I have proposed here is not only timely but urgent. We are at a crossroads in education: There is a crisis mentality accompanied by a flurry of activity to design assessment and accountability systems that both document and promote desired educational change. Current conceptions of reliability and validity in educational measurement constrain the kinds of assessment practices that are likely to find favor, and these in turn constrain educational opportunities for teachers and students. A more hermeneutic approach to assessment would lend theoretical support to new directions in assessment and accountability that honor the purposes and lived experiences of students and the professional, collaborative judgments of teachers. Exploring the dialectic between hermeneutics and psychometrics should provoke and inform a much needed debate among those who develop and use assessments about why particular methods of validity inquiry are privileged and what the effects of that privileging are on the community.

Epilogue

My friend and collaborator, Roberta Herter, who is a veteran English teacher in a large urban school district, tells the story of "Cory," one of her former 10th-grade students. Cory's experience puts a human face on the "detached and impartial" nature of psychometrically sound standardized assessments and illustrates the potential consequences of devaluing more contextualized and dialogic approaches to assessment:

> When he first took the competency exam mandated by the district in 1989, the writing proficiency component required Cory to produce a paragraph of at least five sentences about his experience with friendship. Using the language of the prompt to guide his opening sentence, Cory responded to the test prompt by relating a story about influencing friends to quit smoking while attempting to maintain his relationship with them. The eight sentences he wrote responded to the prompt as if they had been rehearsed, practiced in classroom exercises in preparation for the exam, conforming to the minimum response required to pass the test. The anonymous readers of his exam both rated him a 3.5 on a 5.0 holistic scale, a sufficient score to pass the paragraph portion of the exam.
>
> Cory failed the exam, however, because he did not pass the multiple choice portion of the writing test. Even though his writing demonstrated that he could apply mechanics appropriately, he also needed to demonstrate that he could recognize errors such as misplaced punctuation marks and lack of subject-verb agreement in decontextualized sentences. The decontextualized editing tasks required of the multiple choice portion of the exam failed him.
>
> When contrasted with the lively writing from his folder, work collected over a semester, the writing test distorted and underestimated Cory's capabilities. He wrote on racism, Malcolm X, teen pregnancy, drugs, issues important to him and the community in which he lives. He stood out among his peers as a good writer—a thoughtful, intelligent student who put his writing to real purposes—letters to pen friends, the eulogy he wrote and delivered at his uncle's funeral, and plays on current issues of interest to his classmates read and performed in class.
>
> His versatility as a writer, demonstrated by his ability to write for a variety of audiences in appropriate registers, set him apart from many of his peers who had not achieved Cory's degree of competence. Where Cory's test profile painted a picture of a formulaic writer who could not recognize errors in English usage, his folder showed evidence of a purposeful writer capable of producing controlled and coherent prose. His letters, raps, library reports for science and history, his journal documenting his personal growth and changing attitudes were powerful indicators of his potential for success both in and out of school. In an interview on his own learning he defined education as something ultimately, "you have to do for yourself." He showed himself to be a responsible student, a reflective, critical thinker, conscious of the choices afforded him by the school he attended.

Cory's failure on the exam consigned him to a reading competency class and a class in writing improvement, a low-track English elective where students who fail the exam label themselves LD or learning disabled. Both of these tracked classes were designed to prepare students to pass the test so they might receive an endorsed diploma at graduation—classifying them as having achieved minimum competency in basic skills of reading, writing, and math, or reducing the value of their diploma to a certificate of attendance. But Cory didn't wait to take the test again; he dropped out of day school in his senior year. (adapted from Moss & Herter, 1993)

Notes

[1]Dunbar, Koretz, and Hoover (1991), in a review of empirical research on performance assessment, describe reliability estimates based on the average of coefficients reported for each of nine studies, adjusted via the Spearman Brown formula to reflect an assessment based on a single reader and sample of performance. For the seven studies that took place after 1984, reader reliability ranged from .59 to .91 and task or "score" reliability ranged from .27 to .60. Koretz (1993), describing inter-reader reliability on portfolios from Vermont's statewide assessment, reports that correlations between readers ranged from .33 to .43, with raters assigning the same score between about 50% and 60% of the time.

[2]Other articles have suggested the use of qualitative methods for validity research with less standardized forms of assessment. See, for example, Hipps (1992), Johnston (1992), and Moss et al. (1992). Hermeneutics provides a philosophical underpinning largely consistent with these authors' methodological suggestions. Cherryholmes (1988) suggests that other research discourses, including phenomenology, critical theory, interpretive analytics, and deconstructionism can each contribute, in different ways, to validity research. Mishler (1990) and Johnston (1989) echo similar themes.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: Authors.

Berlak, H., Newmann, F. M., Adams, E., Archbald, D. A., Burgess, T., Raven, J., & Romberg, T. A. (1992). *Toward a new science of educational testing and assessment.* Albany: State University of New York Press.

Bernstein, R. J. (1983). *Beyond objectivism and relativism: Science, hermeneutics, and praxis.* Philadelphia: University of Pennsylvania Press.

Bleicher, J. (1980). *Contemporary hermeneutics: Hermeneutics as method, philosophy, and critique.* London: Routledge and Kegan Paul.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill* (Research Monograph No. 11). New York: College Entrance Examination Board.

Cherryholmes, C. H. (1988). Construct validity and discourses of research. *American Journal of Education, 96,* 421–457.

Cole, N. S., & Moss, P A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201–219). Washington, DC: The American Council on Education and the National Council on Measurement in Education.

Corbett, H. D., & Wilson, B. L. (1991). *Testing, reform, and rebellion.* Norwood, NJ: Ablex.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Holt, Rinehart, & Winston.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? *New directions for testing and measurement: Measuring achievement, progress over a decade,* no. 5 (pp. 99–108). San Francisco: Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test validity.* Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory and public policy.* Urbana: University of Illinois Press.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

Darling-Hammond, L. (1989). Accountability for professional practice. *Teachers College Record, 91,* 59–80.

Darling-Hammond, L., & Snyder, J. (1992). Reframing accountability: Creating learner-centered schools. In A. Lieberman (Ed.), *The changing contexts of teaching* (91st Yearbook of the National Society for the Study of Education). Chicago: University of Chicago Press.

Diesing, P. (1991). *How does social science work? Reflections on practice.* Pittsburgh: University of Pittsburgh Press.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4,* 289–303.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 119–161). New York: Macmillan.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). Washington, DC: The American Council on Education and the National Council on Measurement in Education.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27–32.

Greene, M. (1992). Evaluation and dignity. *Quarterly of the National Writing Project, 14,* 10–13.

Hipps, J. A. (1992, April). *New frameworks for judging alternative assessments.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Johnston, P. (1989). Constructive evaluation and the improvement of teaching and learning. *Teachers College Record, 90,* 509–528.

Johnston, P. H. (1992). *Constructive evaluation of literate activity.* New York: Longman.

Johnston, P. H., Weiss, P., & Afflerbach, P. (1990). *Teachers' evaluation of the teaching and learning in literacy and literature* (Report Series 3.4). Albany: State University of New York at Albany, Center for the Learning and Teaching of Literature.

Koretz, D. (1993). New report on Vermont Portfolio Project documents challenges. *National Council on Measurement in Education Quarterly Newsletter, 1*(4), 1–2.

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1992). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program: Interim report.* Santa Monica, CA: Rand Institute on Education and Training, National Center for Research on Evaluation, Standards, and Student Testing.

Kuhn, T. S. (1986). *The essential tension: Selected studies in scientific tradition and change.* Chicago: The University of Chicago Press.

LeMahieu, P. G., Eresh, J. T., & Wallace, R. C., Jr. (1992). Using student portfolios for public accounting. *The School Administrator, 49*(11), 8–15.

Lieberman, A. (1992). The meaning of scholarly activity and the building of community. *Educational Researcher, 21*(6), 5–12.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 5–21.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11*(1), 3–20.

Messick, S. (1964). Personality measurement and college performance. In *Proceedings of the 1963 Invitational Conference on Testing Problems* (pp. 110–129). Princeton, NJ: Educational Testing Service

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30,* 955–966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012–1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: The American Council on Education and the National Council on Measurement in Education.

Messick, S. (1992, April). *The interplay of evidence and consequences in the validation of performance assessments.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Mishler, E. G. (1990). Validation in inquiry-guided research. *Harvard Educational Review, 60,* 415–442.

Moss, P. A. (1992). Shifting conceptions of validity in educational

measurement: Implications for performance assessment. *Review of Educational Research, 62,* 229–258.
Moss, P. A., Beck, J. S., Ebbs, C., Herter, R., Matson, B., Muchmore, J., Steele, D., & Taylor, C. (1992). Portfolios, accountability, and an interpretive approach to validity. *Educational Measurement: Issues and Practice, 3*(11), 1–11.
Moss, P. A., & Herter, R. (1993). Assessment, accountability, and authority in urban schools. *The Long Term View, 1*(4), 68–75.
Newmann, F. M. (1990). Higher order thinking in teaching social studies: A rationale for the assessment of classroom thoughtfulness. *Journal of Curriculum Studies. 22*(1), 41–56.
Nystrand, M., Cohen, A. S., & Martinez, N. M. (1993). Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment, 1*(1), 53–70.
Ormiston, G. L., & Schrift, A. D. (Eds.). (1990). *The hermeneutic tradition: From Ast to Ricoeur.* Albany: State University of New York Press.
Packer, M. J., & Addison, R. B. (1989). *Entering the circle: Hermeneutic investigation in psychology.* Albany: State University of New York Press.
Perrone, V. (Ed.). (1991). *Expanding student assessment.* Washington, DC: Association for Supervision and Curriculum Development.
Phillips, D. C. (1990). Subjectivity and objectivity: An objective inquiry. In E. W. Eisner & A. Peshkin (Eds.), *Qualitative inquiry in education: The continuing debate* (pp. 19–37). New York: Teachers College Press.
Rabinow, P., & Sullivan, W. M. (Eds.). (1987). *Interpretive social science: A second look.* Berkeley: University of California Press.
Resnick, L. B., & Resnick, D. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford & M. C. O'Connor (Eds.), *Cognitive approaches to assessment.* Boston: Kluwer-Nijhoff.
Ricoeur, P. (1981). The model of the text: Meaningful action considered as a text. In P. Ricoeur (J. B. Thompson, Ed. and Trans.), *Hermeneutics and the human sciences.* New York: Cambridge University Press.
Scriven, M. (1972). Objectivity and subjectivity in educational research. In L. G. Thomas (Ed.), *Philosophical redirection of educational research* (71st Yearbook of the National Society for the Study of Education, Part 1). Chicago: The University of Chicago Press.
Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30,* 215–232.
Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20*(5), 8–11.
Warnke, G. (1987). *Gadamer: Hermeneutics, tradition, and reason.* Stanford, CA: Stanford University Press.
Wiley, D. E., & Haertel, E. H. (in press). Extended assessment tasks: Purposes, definitions, scoring, and accuracy. In R. Mitchell & M. Kane (Eds.), *Implementing performance assessment: Promises, problems, and challenges.* Washington, DC: Pelavin Associates.
Willinsky, J. (1990). *The new literacy: Redefining reading and writing in the schools.* New York: Routledge.
Wolf, D., Bixby, J., Glenn, J., III, & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17,* 31–74.

Received April 19, 1993
Revision received August 10, 1993
Accepted August 30, 1993