Reconceptualizing Validity for Classroom Assessment

Pamela A. Moss, *University of Michigan*

This article explores the shortcomings of conventional validity theory for guiding classroom assessment practice and suggests additional theoretical resources from sociocultural theory and hermeneutics to complement and challenge conventional theory. To illuminate these concerns and possibilities in a concrete context, the author uses her own classroom experience in teaching a qualitative research methods course. The importance of examining cases of assessment practice in context for developing, teaching, and evaluating validity theory is discussed.

Keywords: classroom assessment, hermeneutics, sociocultural theory, validity

This article takes up Sue Brookhart's challenge to conceptualize validity in classroom assessment by thinking "directly about classroom assessment purposes and uses instead of borrowing theory developed in another context" (this issue).[1] I chose to approach this instructive problem by bracketing, for the moment, what I know about how validity theory is conceptualized in psychometrics (American Educational Research Association [AERA], American Psychological Association [APA], & National Council of Measurement in Education [NCME], 1999; Cronbach, 1988, 1989; Kane, 1992, 2001; Messick, 1989, 1994, 1996; Moss, 1992, 1995, 1998; Shepard, 1993, 1997) and how that conceptualization has conventionally been translated to guide classroom assessment (e.g., Airasian, 2001; American Federation of Teachers [AFT], NCME, National Education Association [NEA], 1990; Nitko, 2001; see Black & Wiliam, 1998; Shepard, 2001 for reviews). Rather, I began by reflecting on my own classroom assessment practice in a graduate research methods course and then used that reflection to evaluate the usefulness of conventional validity theory[2] for guiding classroom assessment.

A theory of validity, like any theory, provides us with an intellectual framework or set of conceptual tools that shape both our understanding (We "use it to think with" [Wenger, 1998, p. 58]) and our actions. It illuminates some aspects of social phenomena for consideration and leaves others in the background. As Bernstein notes, an intellectual framework "lend[s] weight to a sense of what are the important issues, the fruitful lines of research to pursue, the proper way of putting the issues" (Bernstein, 1976, p. 41). Or, as the 1999 *Standards* puts it: "it provides a frame of reference to assure that the relevant issues are addressed" (AERA et al., 1999, p. 2). The question I consider in this article is to what extent does our understanding of validity in the measurement profession "assure that the relevant issues are addressed" in classroom assessment and what role might other theoretical perspectives play in providing a more robust validity framework to guide thinking and action?

Although I could certainly frame what I'm doing as a teacher in conventional psychometric validity terms—focusing on particular assessment-based interpretations and their uses; considering the propositions (Kane, 1992), aspects of validity (Messick, 1989), or categories of evidence (AERA et al., 1999) useful or necessary to evaluate the interpretation; and building my validity "argument" (Cronbach, 1988) accordingly—this approach does not provide the most relevant general heuristic for designing and evaluating my assessment practice. While it is, at times, useful and relevant, it does not, I will argue, adequately illuminate or guide me in addressing a number of important issues. Other theoretical resources, both conceptions of validity and of how people learn, that derive from *interpretive* conceptions of social science have proven more broadly useful to me.[3]

In using the case of my own practice to illuminate some limitations of validity theory, it is important to understand that I do not offer it as an example of how things "should" be done, nor do I argue that my classroom experience is directly relevant to other classroom teachers. Each class is both unique and shaped by the institutions of schooling of which it is part (and the frequently routinized practices that occur within them). The institution in which I work differs in many crucial ways from the institutions in which teachers of K–12 students work. And yet, from the perspective of interpretive social science, the usefulness of a case does not depend on its typicality or even on the success of the practices it represents. Rather, it depends on the richness of the detail which allows the readers to draw instructive comparisons with their own contexts of work. As we encounter new cases, our conceptual frameworks expand; we become better able to notice relevant details in subsequent cases (Donmoyer, 1990; Scott, 1998) and better able to act wisely in new situations.

In this article I provide an explanation/illustration of what I believe are the shortcomings of conventional validity theory for helping me design and evaluate my classroom assessment practice. To illustrate these concerns in a concrete context, I use a course that I coteach with my colleague, Lesley Rex: a required introductory course in qualita-

Pamela A. Moss is Associate Professor, 4220 School of Education, University of Michigan, Ann Arbor, MI 48109-1259; e-mail: pamoss@umich.edu. Her areas of specialization are educational assessment, validity theory, and interpretive social science.

tive research methods. In the process, I point to some additional theoretical resources that I have found useful in conceptualizing and evaluating assessment for classroom use. I close with some comments about how we in the measurement profession might expand our capability to be "of use" (Lather, 1999) to classroom teachers and, following Brookhart's suggestion, about the relevance of this sort of reflection for validity theory, more generally.

Assumptions of Validity Theory and Alternative Perspectives

This section outlines some central assumptions of validity theory in psychometrics and provides a brief introduction to the theoretical perspectives that I use to complement/confront them in my practice. Some of these assumptions are explicit in our definitions and extended characterizations of validity research. Some can be arguably inferred from the way validity theory is presented. And some are implicit in our so routinized practices that they shape our conceptions of what validity is.

The following assumptions are considered:

- Assessment is a discrete activity.
- The focus of validity theory is on an assessment-based interpretation and use.
- The unit of analysis for an assessment is the individual.
- Interpretations are constructed by aggregating discrete pieces of evidence to form an interpretable overall score.
- Consequences are an aspect of validity if they can be traced to a source of construct under-representation or construct irrelevant variance (cf. AERA et al., 1999; Messick, 1989).

For each of these assumptions, I will locate them in the literature from which they come, describe the ways in which they do and do not inform my classroom assessment practices, and point to the alternative theoretical resources on which I draw. Here I provide a brief overview of those resources which I will develop and illustrate more fully in the following sections.

In addition to what I have learned from psychometrics, my thinking about my classroom assessment is increasingly informed by a sociocultural perspective on the nature of teaching and learning.[4] Psychometric characterizations of learning—which infer learning from observed changes in individuals' performances over time—have been criticized for viewing learning only as something that takes place "inside the head of the learner" and typically up through a vertical hierarchy of increasingly generalized and abstract knowledge and skills (Beach, 1999; Gee, Hull, & Lankshear, 1996). From a sociocultural perspective, learning is perceived through changing relationships among the learner, the other human participants, and the tools (material and symbolic) available in a given context (Beach, 1999; Chaiklin & Lave, 1993; Cole, 1996; Gee, 1999; Gee et al., 1996; Mehan, 1993; Wertsch, 1998; Wertsch, Del Rio, & Alverez, 1995). Thus learning involves not only acquiring new knowledge and skill, but taking on a new identity and social position within a particular discourse (Gee, 1999; Gee et al., 1996) or community of practice (Wenger, 1998).[5] As Wenger (1998) puts it, learning "changes who we are" (p. 5) "by changing our ability to participate, to belong" (p. 227) and "to experience our life and the world as meaningful" (p. 5).[6] This understanding has important implications for the design of a learning environment, including those aspects that we might distinguish as assessment. (Although we espouse somewhat different theorists, this article shares much in common with Shepard's [2000, 2001] far more extensive discussion of the implications of sociocognitive understandings of learning for classroom assessment; here I foreground the implications this understanding of learning has for the validity of classroom assessment.)

My thinking about validity, especially when making consequential decisions about individuals, is informed by my reading in interpretive social science (e.g., Rabinow & Sullivan, 1987), especially philosophical hermeneutics (Gadamer, 1975, 1987) and the critical dialogue that surrounds it (see Moss, 1994; Moss & Schutz, 2001). Like psychometrics, hermeneutics characterizes a general approach to the interpretation of human products, expressions, or actions. Like psychometrics, hermeneutics provides means of combining information across multiple pieces of evidence and of dealing with disabling biases that readers may bring. Differences between these disciplines lie in the ways in which the information is combined and readers' biases are addressed. Psychometric practices support aggregative strategies for combining information: scores for distinct (ideally independent) pieces of information are (weighted and) aggregated to form an interpretable overall score or grade. Hermeneutics supports a holistic and integrative approach to interpretation of human phenomena, which 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 (Bleicher, 1980; Ormiston & Schrift, 1990; Schmidt, 1995). This iterative process is often referred to as the "hermeneutic circle." Psychometric practices control evaluators' biases by training them to use the same criteria and to focus on the same kinds of evidence. From a hermeneutic perspective, evaluators' disabling biases are addressed, not by trying to train them to "see" things in the same way (which, it could be argued, simply puts a centrally sanctioned bias to work), but by preparing them to let the evidence illuminate and challenge their biases.[7]

Validity Issues in One Case of Classroom Assessment Practice

In the sections that follow, each of the previously stated assumptions about validity theory in psychometrics is considered, in turn, in light of my classroom context and the sociocultural perspective on learning that increasingly informs it. I begin with an overview of the class that I will use to illustrate the relevance and limitations of these assumptions.

The Context of the Class

Qualitative Methods in Educational Research (Moss & Rex, 1997/2001) is intended as a general introduction for doctoral students to various traditions of "qualitative research." Lesley Rex and I developed this class collaboratively and have taught it together four times. The description of the class presented here draws heavily on our shared syllabus. My discussion of the rationale for these choices and the resources that inform them is my own; Lesley would likely frame and value these choices in somewhat different ways as we bring productively different perspectives to the class. Indeed, our differences contribute to the validity of our practices by challenging one another's biases (preconceptions) about teaching, learning, and assessment.[8]

Our goals for the course are to provide: (a) a conceptual overview of a range of research traditions that have been associated with qualitative methods (including ethnography, critical discourse

analysis, feminist poststructural research, and an etic version of qualitative research more congenial to quantitative social science; (b) the opportunity to produce an individual research proposal, reflecting each student's personal research interest, that is similar to what one might write for a small grant; (c) hands-on experience with particular qualitative methods each practiced from within the perspective of one of the research traditions studied; and (d) the opportunity to engage in ongoing critical reflection about different research practices and perspectives that highlight the validity, ethics, and consequences of the choices we make as researchers.

Typically, 24 to 30 students come to the class with quite different career trajectories and years of progress along that path. While the majority of students in the class are first-year students, a substantial minority are typically second-, third-, and even fourth-year students. They come from different programs within and outside the school of education with substantially different epistemological orientations. Most of the students have completed at least one semester of introductory quantitative methods; some have taken one or two additional courses. Some intend qualitative methods to become a central aspect of their research; others are taking the course because it is required. We want the class to be of use to as many students as possible, to move them along their career trajectories in a way that negotiates the delicate balance between learning qualitative methods and honoring their own purposes and interests.

Each of the following sections begins with an explanation of the source of the "assumption" from conventional validity theory, offers an alternative perspective from sociocultural and/or interpretive theory, and illustrates both by drawing on our experience in the qualitative research methods class. Consistent with my argument for the value of extended, richly contextualized cases of practice, it is my hope that readers who work in different contexts will find the issues explored in each section relevant even if they find a different resolution for them.

Conception of Assessment: "Assessment is a discrete activity"

Our very discipline, educational measurement, is predicated on the assumption that assessment can and should be considered as a discrete aspect of the context in which it is used—in this case, the context of learning and teaching. We write standards, textbooks, scholarly books and articles, and so on that bring the practice of assessment to the foreground. We discuss assessment as a distinct phase in the teaching and learning process: the Standards of Teacher Competence characterize assessment activities as occurring prior to, during, and after instruction (AFT et al., 1990); calls for integration of assessment and instruction conceptualize assessment as one component in a cycle of teaching, learning, and assessment (e.g., Airasian, 2001; Nitko, 2001). The conception of assessment as distinct from teaching and learning has consequences for the design of a learning environment. While from time to time I bring "assessment" to the foreground as a discrete issue, I find it is artificial for me to separate out particular activities as "assessment" when I design a learning environment and put it into motion.

When I design a course, I think about the kinds of "*practices*" that are important in the research "*communities*" for which I am preparing students to participate, the kinds of experiences they are likely to bring with them to the class, and the kinds of experiences I want them to have in class to provide resources for their learning (see Wenger [1998] on developing the concept of communities of practice initially presented in Lave and Wenger [1991]). For Wenger, "the concept of *practice* connotes doing, but not just doing in and of itself. It is doing in a historical and social context that gives structure and meaning to what we do" (Wenger, 1998, p. 47).

In order to learn, Wenger argues, we need "opportunities to contribute actively to the practices of communities that we value and that value us, to integrate their enterprises into our understanding of the world, and to make creative use of their respective repertoires" (Wenger, p. 227). A "*community,*" for Wenger, is "the social configurations in which our enterprises are defined as worth pursuing and our participation is recognizable as competence" (Wenger, p. 5). In Wenger's terms, there are at least two communities of practice that are important to consider: the classroom community where learning is the enterprise of the community and the community(ies) of researchers for which we are preparing students to become competent members.

In the qualitative research class, there are two major interrelated projects which are designed to engage students in the practices of qualitative research and learning about qualitative research: an individual research proposal and a collaborative research project where students study the process of developing research proposals. Each of these projects is broken down into a series of smaller activities that builds on the previous activities and introduces a new challenge. While these experiences provide rich opportunities for assessment, they are much more than that: they provide the focus around which learning is organized.

The *individual proposal* describes plans for a research project that draws on qualitative research methods located within a well articulated research perspective of their own choosing. We intend it to further each student's personal research agenda. The development of the proposal is scaffolded throughout the course: in the readings, in interim writing assignments, in class/lab time devoted to work within research support groups, and in individual conferences with the instructors. For the *collaborative research project,* students study the processes through which the individual proposals are developed, focusing especially on how individuals work with the resources and constraints they experience within and beyond the class. We have designed a set of interrelated activities that give hands-on experience with four methods of qualitative research: analyzing open-ended surveys; conducting and analyzing interviews; taking, elaborating, and analyzing fieldnotes; and recording, transcribing, and analyzing group discourse, each practiced from within one of the research traditions explored in the course. These activities provide students with the crucial experience of being the subjects of research as well. Students finish the class with a small group presentation regarding what they have learned about how novice researchers develop research proposals.

Certainly these activities serve purposes conventionally associated with assessment—informing instructional decisions, providing opportunities for self-assessment and feedback from others, monitoring learning, holding students accountable, and so on; equally importantly, they also provide students with opportunities to engage in the practice of qualitative research, not just as a

learning exercise, but as an activity that can have purpose and meaning within and beyond the class. The activities generate the need for information that the readings provide and they generate problems that become the focus of class discussion about validity and ethics. The evidence-based conclusions that students develop from the group research project provide additional resources for teaching and learning. In Wenger's terms, they allow students to build "complex social relationships around meaningful activities . . . in which taking charge of learning becomes the enterprise of a community" (Wenger, p. 272). And they engage the classroom community "in activities that [can] have consequences beyond their boundaries, so that students may learn what it takes to become effective in the world" (Wenger, p. 274). Thus, in this class at least, assessment becomes not so much a discrete set of activities, but rather a way of looking at the evidence available from the learning activities that focus students' "practice" as learners and researchers.

Focus of Validity: "The focus of validity theory is on an assessment based interpretation and use"

Conventionally, validity is conceptualized as referring to an inference or interpretation, and a use or action based on a test score. The 1999 *Standards* defines validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). As the authors of the standards assert: "a rationale should be presented for *each* recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation" (p. 17, italics mine). Similarly Messick, defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (1989, p. 13).[9] Thus the validity argument (or judgment) focuses on ***an*** interpretation or action based on ***an*** instrument. While this focus is at times relevant and useful, it is both too small and too large for most of the decisions I need to make.

For much of what I do, I have no need to draw and warrant fixed interpretations of students' capabilities; rather, it is my job to help them make those interpretations obsolete. What I need to do is make decisions—moment-to-moment, day-to-day, course-to-course—that help students learn, as individuals and as members of learning communities and to study the effects of my decisions (see McMillan, this issue, for a similar argument in the K–12 context). Yes, there are times when a particular interpretation is sufficiently consequential that I need to study its validity more systematically. Even here, however, evidence supporting a consequential interpretation/decision draws from multiple, varied sources (rarely from a single "instrument") about students' learning and the context supporting it. Subsequent sections describe the kinds of evidence that supports (and challenges) these less and more formal types of "assessment."

As mentioned previously, when I first design a class, I think in terms of the overall shape of the activities in the course rather than about single assessment instruments and the particular interpretation I can draw from them. And so, the focus on a single assessment activity is in this way too small to inform what I do. The validity of any particular assessment practice has to do with how it fits with the other assessment practices, in progression, to support (and illuminate) learning. And while each of these activities provides opportunities for us and the students to assess progress, each also serves the multiple purposes that I described.

In one very real sense, assessment is always ongoing: any time I interact with my students is an opportunity for me (and them) to learn about what they are learning and about the quality and effects of my own teaching. In fact, every move I make as a teacher responds to or anticipates a move by students, whether consciously or not (Heritage, 1984; Rawls, 2001;). Even if I focus on the major assignments that students are required to turn in (four different components of the collaborative research project and four drafts of the individual proposal), the interpretations I draw are (hopefully) ephemeral and intended to inform next steps for them and for me.

My goal is to interact with students about their work in a way that supports their learning. This occurs in different ways for different assignments. For the components of the collaborative research project, we give relatively detailed written feedback, much like an editor would, inserted into lines of a draft and write summary comments that highlight what we perceive as strengths and areas in need of work. I keep a working list of issues that I want to be sure to address, but the feedback is not otherwise standardized; it responds to the unique features of each paper. If the class were considerably larger, I would likely use the tools of psychometrics to develop an analytic rubric for feedback on these assignments, but I would consider this a fallback—a necessary efficiency—rather than a better way to evaluate their work. For the research proposal, we meet with the students one on one (except for the final draft where written feedback is given in the manner described above). Here, we find the "feedback" most needed is a conversation that helps draw out what they want to accomplish, that points them toward resources tailored to their interests, that helps brainstorm ways to accomplish their goals or imagine new goals, and/or that suggests ways to manage what seems like an overwhelming task. We keep written notes from these conversations to inform the next meeting. While we intend the feedback to promote rigorous standards of high-quality work, we do not give a grade or otherwise rank the papers into ordinal categories (more on this later). We also ask students to respond to one another's assignments in groups, usually with a series of questions as prompts, to give one another feedback. While I might be able to develop sounder interpretations, in a situation in which time is limited and evidence is continually available, it would be a relatively unproductive use of time to maximize the quality of any single interpretation.

Unit of Analysis: "The unit of analysis for an assessment is the individual"

The methods of educational measurement are most typically used to develop interpretations that characterize individuals, or rather, classes of individuals with the same scores. Group level interpretations are typically based on aggregates of individuals' scores. While other units of analysis could certainly be used, the most common unit of analysis is the individual. Whatever the unit of analysis, the associated scores must be comparable across units and contexts in which the assessment is used. This leads us to standardized assessment. To enhance comparability of scores, we attempt to control the influence of context

through standardization—in essence, we attempt to *fix* the context—so that each individual experiences essentially the same test and contexts of administration.[10] Thus context is treated as separable from inferences about the individual. And we typically make untested assumptions about the generalizability of students' performance from the testing situation to other contexts (Danziger, 1990). As Messick notes, "It is important to recognize that the *intrusion of context* raises issues more of generalizability than of interpretive validity. It gives testimony more to the ubiquity of interactions than to the fragility of score meaning" (Messick, 1989, pp. 14–15, italics mine). This emphasis on individual scores masks the complex role of the social context in shaping those scores and the interpretations they entail.

For the work students undertake in our classroom, the sort of standardization that enables comparability of scores is not feasible, nor would it be pedagogically sound to alter the tasks to make it so. As Wenger suggests:

> One problem of the traditional classroom format is that it both too disconnected from the world and too uniform to support meaningful forms of identification. It offers unusually little texture to negotiate identities: a teacher sticking out and a flat group of students all learning the same thing at the same time. Competence, thus stripped of its social complexity, means pleasing the teacher, raising your hands first, getting good grades. (p. 269)

Such standardization is not consistent with what students would experience doing qualitative work outside this class as competent members of a community of qualitative researchers. It does not allow students to build "complex social relationships around meaningful activities" (Wenger, p. 272).

Even if we focus on the formal assignments in the qualitative methods class, students' experience of them is necessarily and appropriately complex, varied, and partially unique. While students' performances are shaped by the written descriptions of these assignments, which all students see, they are also shaped by a myriad of factors, including the choices students make (e.g., about how to focus their research proposal), by the always partially unique data they encounter in the collaborative project, by the particular class readings and the readings students locate on their own; by their ongoing interactions about their work with other students in the class, with us as instructors, and with their advisors (who often become actively involved in the research proposal); and by more formal feedback from the instructors which is itself tailored to the unique features of each paper. And, of course it is shaped by students' interpretations of the task which, in turn, is shaped by all the perspectives and practices that students bring with them to class from their own experiences outside of class. A change in any one of these features may well affect (a little or a lot) the nature of the "assessment" for a single student, a group of students, or the entire class. In order to interpret and/or evaluate a student's performance, I need to understand the influence of the contexts in which it was produced and to understand the factors that shape that performance.

Consistent with a sociocultural perspective, the most appropriate unit of analysis is the *social situation* (Mehan, 1998)—which entails the recursive relationship between person and context (including the actions of other people, available resources, and larger social structures in which they are encountered)—and claims about individuals must be grounded in *interaction* (see also Wertsch et al., 1995). Understanding students' performance in context is also crucial to enhancing fairness with more and less standardized assessments. As Mehan notes,

> By moving beyond the states and traits of individuals to social situations as the unit of analysis, it does not blame low achieving students' school difficulties on their lack of motivation, diminished linguistic skills, or deficient cognitive styles. . . . [Students' performances can be] recast as collaboratively constructed and continuously embedded in face-to-face interaction in social environments. (pp. 251, 254)[11]

From this perspective, the most useful sort of evidence is that which documents the interaction—the ongoing effects of actions on other actions. This can and should involve many different levels of analysis, including interaction at the microlevel of moment-to-moment interaction as well as at larger grain sizes such as revisions to assignments in response to instructors' feedback, or, even larger, changes in the nature of students work across cohorts experiencing somewhat different course design features. These can certainly be combined to develop an interpretation that focuses on the individual; these can also be combined to focus on the effects of the particular instructional activities. While much of the time, the evidence supporting my instructional decisions is examined only informally, from time to time, it would be useful to engage in more disciplined study of my practice and its effects. This requires explicit documentation of the interaction so that it can be reviewed with various questions and associated analytical lenses in mind. It might include examining videotapes or audiotapes of classroom discourse or interactions with individual students. It might involve case studies of individual students, examining changes in their performance across time in light of the interactions that shaped them. It might focus on the effect of a particular practice or activity, where I examine student work in response to variations in that activity, and interview students about how they interpreted and used the resources available to them. The growing tradition of action research in teaching is rich with relevant examples (e.g., Cochran-Smith & Lytle, 1993; Power & Hubbard, 1999). Teaching portfolios, of the sort I am required to compile for routine reviews, can prompt this sort of systematic analysis.[12] Approaches to enhancing and documenting the validity of these sorts of interpretations, which combine quite disparate sources of evidence, is discussed in the following section.

Combining Evidence: "Interpretations are constructed by aggregating judgments from discrete pieces of evidence to form an interpretable overall score"

Having multiple sources/pieces of evidence to inform a consequential interpretation/decision is a fundamental feature of the epistemology and ethics in any of the social science perspectives that I have encountered. Similarly, although it is framed in different ways in different social science disciplines, illuminating and challenging (disabling) biases is also fundamental. The practices of educational measurement are flush with techniques for aggregating evidence to an overall score, with associated standard error(s), from which interpretations/decisions can be made; however, they have very little to offer when aggregation is not possible or desirable. Thus while certain types and combinations of evidence lend them-

selves well to measurement perspectives, others do not. Aggregation entails that (at least categorical) judgments be made about discrete pieces of information so that once an assessment system is developed, these judgments can be algorithmically combined (weighted and accumulated) to form a "score" that has a predetermined interpretation associated with it.[13] The very definition of validity in the testing *Standards* associates it with test scores: "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA et al., 1999, p. 9). The AFT, NCME, and NEA *Standards for Teacher Competence in Educational Assessment of Students* asserts that "teachers should be skilled in administering, scoring and interpreting the results of . . . teacher produced assessment methods" which includes "being able to use guides for scoring essay questions and projects, stencils for scoring response choice questions, and scales for rating performance assessments . . . [to] produce consistent results" (p. 4). That assessment items are "scored" (i.e., result in at least categorical judgments) seems to be taken for granted. I have at least two concerns about this assumption and the methodological advice associated with it: (a) it is inadequate to inform a large part of my practice as a teacher and (b) it risks shaping my practice to conform to its vision of assessment. Where conventional psychometric tools fit, I would certainly use them; where they do not, I need other sorts of advice. Ironically, the testing standards also assert (more weakly than some of us would prefer):

> In educational settings, a decision or characterization that will have major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision. (p. 146)[14]

Citing the example of identifying students with special needs, the authors of the *Standards* note: "it is important, that in addition to test scores, other relevant information (e.g., school record, classroom observation, parent report) is taken into account by the professionals making the decision" (p. 147). And yet, psychometrics has little advice to offer about how to combine such evidence into a well warranted interpretation or decision—a task to which interpretive social science is well suited.

In the qualitative methods class, there are many sources of evidence and conceptual lenses for analyzing them that could inform my interpretation of each student's practice and trajectory toward becoming a competent researcher (and my own progress toward becoming a better teacher). These include not only what students have produced in response to the multiple formal assignments, but how they engaged the tasks as evidenced through successive drafts, ongoing conversations, and observations of their interactions with others, and my knowledge of what resources they had available to them to support this work. The appropriateness of using these different sources of evidence depends on the purposes to which the interpretations are put.

As suggested previously, most of the "interpretations" I draw are part of the ongoing logic of practice (Bourdieu, 1990), continually updated as teaching and learning unfold; some, however, are more formal, resulting in extended written or oral prose and requiring more explicit consideration of their validity. These most typically include responses to single assignments as previously described, conversations with students about their progress in the course, and letters of recommendation (which may incorporate evidence from other courses or work relationships with the student). Some involve more consequential decisions. And, the more consequential the interpretation, the more crucial the explicit and rigorous consideration of its validity. This is especially true when an interpretation, like a letter of recommendation, is to be used outside the class to represent my perspective on the student's work to others or when a consequential decision (e.g., satisfactory/unsatisfactory) about a student must be made. (Smith's [this issue] conception of sufficiency of information is useful here as long as it is considered in light of the purpose that the assessment is intended to serve and expanded beyond quantitative indicators to include multiple means of evaluating sufficiency [a point Smith also makes]). Here, I turn to hermeneutics and to the approaches to validity consistent with the research practices that we teach in this qualitative class.

In the introduction, I offer a brief overview of hermeneutics as a theoretical resource for an interpretive approach to validity. In addition, readers of this article will find many "qualitative" methods texts (e.g., Erickson, 1986; Patton, 2001) that provide practical advice that is consistent with that theoretical perspective. Let me summarize some common features of validity practice for these sorts of extended prose interpretations.[15] The description represents a more rigorous approach that would likely be relaxed depending on the consequences associated with the interpretation. Clearly, choices made in the design of a class limit the evidence one has available; with extended prose interpretations, however, one can develop interpretations that are appropriate to the evidence and can situate the interpretation within descriptions of the evidence on which they are based. Interpretations are formed and refined through an iterative process of repeatedly testing them against the available evidence, until each piece of evidence can be accounted for in a coherent interpretation of the collection of available evidence. Having multiple sources of evidence gathered across time and situation enhances the validity of an interpretation. If I do not have enough evidence to address an issue that I believe needs to be addressed, I can seek additional evidence. The vigorous attempt to discover problems with the proposed interpretation—the search for disconfirmatory evidence and for alternative interpretations that account for the same evidence—is central to the development of well warranted interpretations. The interpretations and supporting evidence are best presented in such a way that the reader (or listener), including the student, can evaluate the interpretations and supporting evidence for himself or herself and be allowed to offer counter interpretations and counter examples. As Thompson (1990) notes, valid interpretations are justified, not imposed, on the person(s) about whom they are made. Having multiple readers (interpreters) contributes to the strength of the validity of the interpretation. This can involve an independent evaluation of the same evidence or an audit of the trail of evidence leading to the conclusions. While agreement among readers may be a welcome outcome, disagreement is also a validity resource (Moss & Schutz, 2001). Lesley and I, for instance, sometimes disagree about the features of a strong paper—in part due to differences in our disciplinary perspectives and in part to our own

unique styles. Validity is enhanced because we are each provoked to see and possibly reconsider criteria that we have taken for granted, even if we still retain different readings of a paper. While we could force ourselves to use the same criteria for evaluation, our students would not necessarily be better served: our differences reflect real differences among thoughtful proponents of qualitative research and students will need to learn to negotiate these differences for themselves. The goal for us as teachers is to make those criteria and the rationale underlying them explicit in our interactions with students.

Role of Consequences: "Consequences are an aspect of validity only if they can be traced to a source of construct underrepresentation or construct irrelevant variance"

The relationship between validity and consequences remains controversial within the field of educational measurement. The perspective characterized above, which sees a limited role for consequences in the understanding of validity, can be traced to Messick (1989) and the testing *Standards* (AERA, APA, & NCME, 1999). This circumscribed perspective allows us to consider the soundness of an interpretation absent any effects of its production and reception that cannot be traced to problems with the interpretation itself. Others see a larger role for consequences in their conception of validity (e.g., Cronbach, 1988; Shepard, 1993, 1997; see Moss, 1998, for a brief review), focusing on whether the test serves the purpose it was intended to serve and on unintended (positive and negative) consequences.

Whatever one's definition of validity, with classroom assessment, understanding these effects is crucial to sound practice. I might go so far as to argue that validity in classroom assessment—where the focus is on enhancing students' learning—is primarily about consequences. Assuming interpretations are intended to inform instructional decisions and that instructional decisions entail interpretations about students' learning, it is on evidence of their (immediate, long-range, and cumulative) effects on which their validity primarily rests. The insistence on interactional evidence at various grain sizes—for understanding an action in the context in which it was produced and received—also highlights the central role of consequences in validity.

A sociocultural understanding of learning prompts a wider consideration of the nature of consequences than is typically considered in routine assessment practice. The success of the class, in Wenger's terms, depends on the extent to which we move students along the trajectory toward becoming competent participants in a community of qualitative researchers or competent users of qualitative research in the community in which they want to participate. This certainly includes what they know and can do—the conventional foci for assessment—but it also includes how they are positioned (to relate to that knowledge and to the other people in their social contexts) and who they are, as a result, *becoming* (see, e.g., Gee et al., 1996; Greeno, in press; Lather, 1991; Wenger, 1998). As previously illustrated, this understanding has major implications for the development of class activities. It also illuminates the importance of bringing an additional set of questions to the evidence we gather about our classroom practices and students' response to it. I have found useful tools for asking questions about identity and positioning in the work of critical discourse analysts Gee (1996, 1999) and Fairclough (1989, 1995). When I examine the effects of my pedagogy (including my assessment practice), I seek evidence not only about the soundness of the knowledge I'm enabling students to construct, but about what kind of selves I am helping them become. What am I doing that shapes their sense of what it means to be a teacher or a researcher or a learner that I may not even be aware of? One strategy I found particularly useful is to examine my practice from different perspectives. I'm starting to ask myself questions like, how do I interact with my colleagues in ways that are different from the ways I interact with my students? Why? Assuming my goal is to position students to become competent members of a community of researchers, how do those differences enhance or detract from that goal? For instance, when I talk with students or colleagues about their work, how do I position myself with respect to them differently? What authority do I invite them to take in our dialogue about their work? When I make a comment about a student's paper, am I pushing the student to defer to my judgment or am I inviting him or her to use my evaluation to reach his or her own well warranted judgment which may (for good reasons) disagree with mine? The presence of a co-teacher with whom I do not always agree on what constitutes sound work is a productive "design" feature of the learning environment that invites students to engage thoughtfully in these discussions to reach their own working conclusions (just as they would, for instance, if they were responding to reviews of a manuscript submitted for publication). I am not saying that I always conclude that I should interact with students as I do with colleagues. Rather, I'm saying that the comparison helps highlight ways of talking and acting that I take for granted so that I can self-consciously consider them.

Informal consideration of interactional evidence with these sorts of questions in mind helped me make the decision to abandon grades, whenever possible. I had always found the giving of grades to require a substantial commitment of time to develop a meaningful rubric and assign scores fairly—time that took me away from tasks that seemed to have a higher pedagogical value. I began to attend more explicitly to how they shaped my interactions with students about their work, both before and after the assignment of the grade. Conversations too frequently focused on what I wanted, on what I (could specify I) considered necessary for an A, or on why a higher grade than the one I had assigned was fair. When I gave students opportunities to revise their work to improve the grade or I postponed the giving of a grade until revised versions were turned in, I found the revision typically accomplished just what I had asked for and nothing more. Ungraded rubrics functioned in much the same fashion. As Shepard (2003) notes: "competitive grading practices seem to be so pervasive in U.S. classrooms that the purpose of rubrics has been corrupted from criteria that make the features of excellent work accessible into a point system used for defending and quarreling over grades" (p. 176). I don't want the capital (Bourdieu & Wacquant, 1992) in my classroom to be grades or even my approval; it will not sustain students (as professionals) outside the classroom. I want it to be doing something that is meaningful and useful within the context of classroom and the relevant research communities.

Now, in the qualitative methods class, we do not use grades or other scoring rubrics during the term. To deal with the institutional requirement for an end-of-term grade, we treat an "A" as satis-

factory performance. We tell students that if they fully engage with the activities of the class, fulfill their responsibilities to us and the other students, and take advantage of the learning opportunities to improve their practice, they will get an A. We promise to let individuals know if we become concerned about their performance in the class and to give them an opportunity to change the nature of their participation. It happens only rarely, in this class, that a student's work is sufficiently problematic to be considered unsatisfactory. Where it does, we deal with it on a case-by-case basis (and in consultation with the student), considering options of a lower grade, an incomplete, or withdrawal from the class. This actually allows us to give rigorous and critical feedback in an atmosphere where students are not afraid of its consequences. With that practice, I have noticed changes in the nature of the conversations I have with students about their work; we now focus more on what they want to accomplish, on our collaborative critical reading of their progress, and on how that might fit with (and/or challenge) current understandings of their primary research community. That does not mean that students do not challenge me—and indeed I want to encourage them to become the kind of scholars who mount thoughtful challenges when they disagree; it does, however, allow us to focus on improving the work, rather than improving the grade. Although many students have reported that they like this practice (because it allowed them to take risks and not worry about the grade), some students have expressed discomfort, arguing that they need the grades for motivation and accountability. (And, indeed, I have noted it is sometimes hard to compete for students' attention in an institutional economy that privileges grading.) We seek to instill other kinds of accountability that are more consistent with the practices of the communities of teachers and researchers we are preparing them to join: by encouraging students to engage in work they find meaningful and that has a purpose beyond the demonstration of learning and by setting up collaborative participation structures so that students are accountable to one another. Again, this is the kind of "motivation" they will need to sustain them as they move out of our classrooms and into their own professional communities. Of course, not all teachers will have the flexibility to develop a grading policy like ours and that policy will not serve all (classrooms of) students well. But, I would argue that cases that contrast with our own experience can illuminate features of our contexts that we might take for granted as "the way things are." The point I hope readers will take away is the importance of considering (and studying) the role that different grading policies can play in the classroom environment.

Implications for Teaching and Research

In the previous sections, I've illustrated a number of ways in which our conventional conception of validity in educational measurement is not sufficient for my practice as a classroom teacher. Were I to revise my practice to enact more faithfully its conception of good practice, it would partially compromise the environment we have designed as a resource for students' learning. Conventional conceptions of validity privilege standardized forms of assessment. The authors of the testing standards acknowledge this and point to its limitations for less standardized forms of assessment, including classroom assessment.[16] While this representation of validity can provide important guidance for sound practice with standardized assessments, not all forms of classroom assessment can or should be standardized. Standardization has consequences for learning. It enacts a particular relationship among students, teachers, and content—it proposes a particular identity for learners—that is not consistent with my goals as a teacher.

When standardization is not appropriate, other conceptions of validity are needed. Ours is not the only social science discipline that cares deeply about the quality of interpretations and the soundness of the evidence supporting them. The move away from standardization and the validity practices that support it need not result in a trade-off between relevance to classroom practice and validity (or reliability), an issue Smith (this issue) raises. Interpretive practices of validity, comparably rigorous and historically grounded, can support interpretations when standardization is neither feasible nor desirable. Indeed, the testing standards themselves were developed through practices that more closely resemble the interpretive approach to drawing and warranting conclusions than to the conception of validity that psychometrics supports.

Within the measurement community, a number of scholars who study classroom assessment have begun to turn to interpretive approaches to validity for dealing with less standardized forms of assessment and to qualitative or interpretive research more generally for helping us understand how assessment works in particular social contexts. Shepard (2001), for instance, argues: "evaluating open ended tasks and drawing valid inferences from both formal and informal data sources requires new methods of data analysis and interpretation. . . . [They] create a profoundly greater need for teacher judgment and qualitative methods of inquiry" (p. 1088). Similarly Black and Wiliam (1998) call for complementing quantitative measures "with richer qualitative studies of processes and interactions within the classroom" (p. 44). Some scholars privilege interpretive social science as the basis for classroom assessment (e.g., Gipps, 1994, 1999; Johnston, 1989).

Implications for Teaching Classroom Assessment

A similar sort of argument about the limits of a single set of general principles could be made for the way we attempt to develop and teach generalizable knowledge about classroom assessment. Black and Wiliam (1998), in their extensive review of the literature on classroom assessment, explain that the problem with attempting to draw generalizing conclusions about features of formative assessment practices is that they are all always situated in complex social situations which co-determine the results. This resonates with my own experience. Even within the circumscribed social world of my classroom, the usefulness of any given feature of my practice (say how to give feedback on students' papers), varies with the nature of the assignment, its place in the sequence of activities, the particular issue the paper raises, the particular experiences and perspectives a student brings to it, and so on. And, as Mehan notes: "Good educational practice does not exist outside of a particular educational context and 'just good teaching' is not just good teaching at all, but a complex process of combining information from a number of different sources to produce practice well adapted to the population and setting at hand" (Jacob & Jordan, 1993, in Mehan, 1998, p. 257).

Given this perspective, I believe validity needs to be conceptualized, not

just in the "context of classroom assessment" but in the contexts of the classrooms in which the assessment occurs. And, teachers will need to prepare to respond, resourcefully, to the always partially unique features of the learning contexts in which they work. When I have taught courses in assessment to (prospective) teachers, I have designed the course to give participants practice with the kinds of decisions they will likely need to make and the kinds of conversations they should be able to participate in knowledgeably within and beyond the classroom. Further, I have always asked teachers to develop and evaluate plans and learning/assessment activities over the course of the term within the context of a particular unit of instruction so they can consider how the activities work together to support and provide evidence about students' learning. By working within a particular unit of instruction, they can read deeply about instruction and assessment in that area and about relevant cognitive and sociocultural research; they can ask their advisors for resources and feedback relevant to the unit; and they can share what they are learning with their colleagues so that everyone in the class can learn from the different extended cases of assessment practice.

The portfolio they prepare over the course of the term typically includes: an overview of the instructional context and curriculum unit; an evaluation plan and rationale (including both formal and informal assessment); actual assignments and/or detailed descriptions of major learning/assessment activities; case studies of students engaging in work relevant to the unit over time; evaluation of a class set of papers relevant to the unit; an annotated bibliography of references consulted; a summary of validity evidence available/questions remaining to support their interpretations of students' learning; and a critical review of a published standardized test used in the same instructional context. As participants read and respond to one another's work, they expand their repertoire of cases of assessment practice. If I were teaching the course now, I would probably build in an exhibit on analyzing videotapes of classroom discussion relevant to the unit; I would also encourage participants to take a course in action research, where the focus of research would shift subtly from their students' learning to their instructional practices. The required case studies of individual students are particularly useful for having teachers practice careful interpretive work consistent with hermeneutics: reading over multiple samples of their students' work (possibly accompanied by observation/interview notes), developing initial interpretations and trying them out against the existing evidence, looking for counterevidence and trying out alternative interpretations, and revising the interpretations accordingly. While they will not or can not typically approach assessment with that degree of care, it models an interpretive practice they can internalize, streamline as necessary, and use as a touchstone of rigor in thinking about the validity of their interpretive practices. I have taught traditional conceptions of reliability and validity in conjunction with the preparation of the critical review of an externally imposed test used by their district/state. Teachers learn both about how test developers evaluate the quality of their tests and about how they might evaluate the relevance/impact of the test in their own classroom context: by comparing the tested domain to the domain enacted in their own curriculum and, if possible, interpreting classroom level reports in light of the local curriculum and interviewing students' about their experiences. We should not lose sight of the fact that large-scale assessments are also administered and used in local (classroom, school, district) contexts and that a rich understanding of students' performances requires an understanding of that context. Following Mehan (1998), I believe courses in assessment need to promote "a new relationship between teacher, researcher and pedagogical knowledge. . . . In this new configuration, the teacher moves from being a passive recipient of packaged research knowledge to a collaborative constructor of pedagogical knowledge useful in local circumstances." (p. 258).

The Role of Cases in Validity Theory

The most basic question underlying this article is what role cases of assessment practice should play in the development and/or representation of validity theory and assessment pedagogy? A straightforward answer is that the principles are necessarily general and that we need cases to illustrate how they can be instantiated in practice. Shepard (1993), for instance, in her article on validity includes a series of extended cases of validity research that encompass the perspectives of different scholars "to illustrate . . . how a set of essential validity questions might be outlined" (p. 430). "These kinds of evaluative investigations motivated by claims and counterclaims reflect the principles of construct validation in its fullest sense" (p. 436).[17] While these uses of extended cases play an important role in our understanding of validity, well chosen cases can provide more than illustrations of sound practice: they can provide crucial opportunities to critique and revise the general principles and, where necessary, to create or appropriate new ones. This is particularly true (a) if at least some of the cases we examine have not already been fully shaped or "colonized" by our general principles; and (b) if we are willing to undertake the effort with what has been called a "hermeneutic attitude": the willingness to risk our own preconceptions and the belief that we have something to learn from the other (Bernstein, 1985; Gadamer, 1987).

But, how does knowledge advance from the in-depth study of single cases? Here, a different conception of "generalizability"—both of validity principles and of knowledge about classroom assessment—may be useful to consider: one that goes beyond the straightforward applications of generalized propositions to concrete situations. Expertise in ill-structured domains (like designing and evaluating assessments) does not develop only, or even primarily, through the acquisition of abstract concepts that can then be routinely applied (Beach, 1999; Bransford & Schwartz, 1999); rather, it develops through concrete experiences that allow us to develop increasingly sophisticated capabilities to respond to (learn from) the always partially unique features that each case represents. Bransford and Schwartz distinguish two conceptions of transfer: (1) "direct application" which "asks whether people can apply something they have learned to a new problem or situation" (p. 67) and (2) "preparation for future learning" which "broadens the conception of transfer by . . . shift[ing] the focus to assessments of people's abilities to learn in knowledge-rich environments." As they note, "when organizations hire new employees, they do not expect them to have learned everything they need for successful adaptation. They want people who can learn, and they expect them to make use of resources (e.g., texts, computer programs colleagues) to facilitate this learning"

(p. 68). "The learning experiences 'set the stage' for further noticing, and their effects cannot be reduced to the mere replications of a particular experience per se" (p. 74).

In the context of classroom assessment and of validity research more generally, we need to develop and maintain a rich repertoire of cases: not just those that illustrate how our guiding principles can be thoughtfully applied but, equally important, those that have not already been shaped by our principles so that we can learn about their limitations. With classroom assessment, I think of examples like Lampert's (2001) year-long analysis of her problem-based mathematics teaching with fifth graders, Rex's collaborative interactional ethnographies of secondary English classrooms (e.g., Rex, 2001, Rex & McEachen, 1999), Lee's analysis of her cultural modeling approach to pedagogy with secondary English students (e.g., Lee, 2001), or Rogoff's and colleagues' (Rogoff, Turkanis, & Bartlett, 2001) representation of their work in a school-wide learning community. These cases provide us with vicarious experiences of how successful teachers create learning environments and evaluate their students' work using evidence based in interaction. Moreover, they position us (productively), not as the experts whose role it is to reshape these environments in our own images of what constitutes good assessment, but as fellow learners who can think with teachers and teacher educators about how to conceptualize validity to be of use in particular contexts of assessment.

Any vital epistemology needs to contain within itself the tools to call even its most fundamental principles into question. Messick (1989) was well aware of this when he advised us to invite study of the practices of educational measurement from alternative disciplinary perspectives to illuminate the technical and value assumptions underlying each.[18] The importance of encountering outside perspectives to illuminate what "we" take for granted (as natural, normal, the "way things are done") and to provoke critical, action-orienting, self-reflection is a theme that resonates across multiple philosophies of social science and logics of inquiry (e.g., Bernstein, 1985; Cole, 1996; Gadamer, 1987; Hoy & McCarthy, 1994; Messick, 1989). Careful consideration of concrete cases and of alternative conceptions of validity, taken together with the willingness to risk our own preconceptions, can only strengthen the epistemological moorings of our profession and our ability to be of use to others.

Notes

This article was written while I was collaborating with Jim Gee, Ed Haertel, and Diana Pullin on an article entitled "The Idea of Testing: Expanding the Foundations of Educational Measurement" (Moss, Pullin, Gee, & Haertel, 2002). My thinking about classroom assessment has been enhanced by my conversations with them and affinities will be found between the two articles. My work in validity theory has been supported by grants from the Spencer Foundation and the National Academy of Education. I am grateful to Ed Haertel, Laura Haniford, Renee Miller, Lesley Rex, and Mark Wilson for comments on an earlier draft of this article.

[1]Taylor and Nolen (1996) make a similar argument.

[2]By "conventional validity theory," I refer primarily to the way in which validity is characterized in the 1999 *Standards for Educational and Psychological Testing,* although many of the assumptions I highlight can be traced to the work of other scholars.

[3]"Current conceptions of validity in educational measurement have evolved from a 'naturalist' view of social science, which 'maintains that the social sciences should approach the study of social phenomena in the same ways that the natural sciences have approached the study of natural phenomena' (Martin & McIntyre, 1994, pp. xv–xvi). From this perspective, primary goals of social science are nomological or generalizable explanation or prediction. An alternative view of social science, frequently labeled 'interpretive' (e.g., Bohman, Hiley, & Shusterman, 1991; Rabinow & Sullivan, 1987), sees the subject matter of the social sciences as fundamentally different from the natural sciences and the methods and aims of natural science as inadequate to represent social phenomena. Traditions as diverse as ethnography, hermeneutics, phenomenology, critical theory, and postmodernism (any one of which also comprises diverse perspectives) have been located within this conception of social science. One primary goal that interpretive traditions share is to understand meaning in context." (Moss, 1996, p. 21).

[4]While some limit the term "sociocultural" to research that derives from the work of Vygotsky, others use the term more broadly to refer to a constellation of perspectives that attend to the dialectical relationship between social structure and local practice of individuals in context, which is the perspective I use here.

[5]Adapted from Moss et al., 2002.

[6]Wenger (1998) illustrates this perspective with an ethnographic study of how people learn to become medical claims processors for an insurance company. In their earlier work, Lave and Wenger (1991) conducted case studies of quite diverse communities of practice (see also Lave, 1988). Similarly, Beach (1999) and Gee and colleagues (1996) note instructive contrasts between the kinds of learning that typically occur in school with those in various work places.

[7]I add the modifier "disabling" since from some social science perspectives (including my own), biases (or preconceptions) are seen as inevitable and, in fact, essential for making meaning. Bernstein (1985), following Gadamer (e.g., 1975) argues that there is no knowledge without foreknowledge—without preconceptions or prejudices. "The task is not to remove all such preconceptions, but to test them critically in the course of inquiry. . . . to make the all important distinction between blind prejudices and 'justified . . . [or enabling] prejudices that are productive of knowledge' " (p. 128).

[8]The term "qualitative" is one we inherited with the title of the course. It characterizes many methods that can be associated with a wide variety of research perspectives from hypothesis testing experimental research to postmodern studies. As such, it risks being theoretically incoherent. Many practitioners of traditions that would be considered "qualitative" (including those we explore in this course) use other terms that are more specific to their traditions. When a broad term is needed, I prefer "interpretive," as described in note 4.

[9]Note the difference in the role of the word "use" or "action" in these two definitions, which themselves have somewhat different implications for the nature of validity research.

[10]Although new directions in cognitive psychology and psychometrics (Hattie, Jaeger, & Bond, 1999; Mislevy, Almond, & Steinberg, 2003; National Research Council, 2001; Pellegrino, Baxter, & Glaser, 1999) promise more nuanced characterizations of individuals and situations, the methods still require a fixed set of possibilities for operational uses of assessments.

[11]Cazden (2001) for instance cites evidence that shows how teachers interact differently with students perceived to be of different capabilities and how this, in turn, gives them differential access to knowledge (see also, Moss et al., 2003).

[12]Teaching portfolios that organizations such as INTASC and NBPTS are developing for professional development and certification, and collaborative study groups such as the Brookline Teacher Research Seminar (Cazden, 2001), the Santa Barbara and Michigan Discourse Analysis Groups (1992; www.owp.soe.umich.edu/McDiG), or the Henry Ford study group with whom I worked (Clark et al., 1996), provide systematic opportunities for reflecting on evidence about the relationship between teaching and learning.

[13]Porter (1995), a historian of statistics and quantitative reasoning, in fact argues that this approach to making decisions, which rests its warrant in faithfully following sanctioned methods, essentially relieves decision makers of personal responsibility for the decision.

[14]Of course, the import of this statement depends on what your conception of validity is.

[15]Some theorists operating within an interpretive social science perspective eschew any a priori conception of validity, arguing that the conception of validity should arise from the situation. See, for instance, Lather's (2001) "situative" approach to validity. My own sense is that general principles and concrete cases work best in dialectical relationship.

[16]"In all cases, however, tests standardize the process by which test-taker responses to test materials are evaluated and scored. . . . Although the *Standards* may be applied most directly to standardized measures generally recognized as 'tests,' . . . it may also be usefully applied in varying degrees to a broad range of less formal assessment techniques. Admittedly, it will generally not be possible to apply the *Standards* rigorously to unstandardized questionnaires or to the broad range of unstructured behavior samples used in some forms of clinic- and school-based psychological assessment, . . . and to instructor-made tests that are used to evaluate student performance in education and training." (AERA, APA, & NCME, 1999, p. 3)

[17]Gadamer sees the relationship between principles and cases as more than one of illustration. He highlights the importance of concrete cases in understanding the very meaning of norms or laws: "What one considers the right decision determines the standard itself" (Gadamer, 1975/1994, p. 570). In fact, norms are "indeterminable without the concrete situation in which one thing is preferred to another" (1981, p. 92). Thus, for Gadamer (1981) the set of interpreted concrete cases to which a norm has been applied is at least as important as the norm itself: "the body of precedents (the decisions already laid down) is more crucial for the legal system than the universal laws in accord with which the decisions are made" (p. 82).

[18]"A Singerian inquiring system starts with the set of other inquiring systems . . . and applies any system recursively to another system, including itself. The intent is to elucidate the distinctive technical and value assumptions underlying each system application and to integrate the scientific and ethical implications of the inquiry." (Messick, 1989, p. 32)

References

Airasian, P. (2001). *Classroom assessment: Concepts and applications* (4th ed.). Columbus, OH: McGraw-Hill.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *Standards for teacher competence.* Lincoln, NB: Buros Institute. Available at: www.unl.edu/buros/article3.html

Beach, K. (1999). Consequential transitions: A sociocultural expedition. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (Vol. 24, pp. 101–140). Washington, DC: American Educational Research Association.

Bernstein, R. J. (1976). *The restructuring of social and political theory.* Philadelphia: The University of Pennsylvania Press.

Bernstein, R. J. (1985). *Beyond objectivism and relativism: Science, hermeneutics, and praxis.* Philadelphia: The University of Pennsylvania Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5,* 7–14.

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

Bohman, J. F., Hiley, D. R., & Shusterman, R. (1991). The interpretive turn. In D. Hiley, J. F. Bohman, & R. Shusterman (Eds.), *The interpretive turn* (pp. 1–16). Ithaca, NY: Cornell University Press.

Bourdieu, P. (1990). *Logic of practice* (R. Nice, Trans.). Stanford, CA: Stanford University Press.

Bourdieu, P., & Wacquant, L. J. D. (1992). *An invitation to reflexive sociology.* Chicago: University of Chicago Press.

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences.* Cambridge, MA: MIT Press.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education,* 24, 61–100.

Cazden, C. B. (2001). *Classroom discourse: The language of teaching and learning.* Portsmouth, NH: Heinemann.

Chaiklin, S., & Lave, J. (Eds.). (1993). *Understanding practice: Perspectives on activity and context.* Cambridge, UK: Cambridge University Press.

Clark, C. T., Moss, P. A., Goering, S., Herter, R., Lamar, B., Leonard, D., et al. (1996). Collaboration as dialogue: Teachers and researchers engaged in conversation and professional development. *American Educational Research Journal, 33*(1), 193–231.

Cochran-Smith, M., & Lytle, S. L. (1993). *Inside/outside: teacher research and knowledge.* New York: Teachers College Press.

Cole, M. (1996). *Cultural psychology: A once a future discipline.* Cambridge, MA: The Belknap Press of Harvard University Press.

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

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory and public policy* (pp. 147–171). Urbana: University of Illinois Press.

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research.* Cambridge, UK: Cambridge University Press.

Donmoyer, R. (1990). Generalizability and the single-case study. In E. W. Eisner & A. Peshkin (Eds.), *Qualitative inquiry in education: The continuing debate* (pp. 175–200). New York: Teachers College Press.

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.

Fairclough, N. (1989). *Language and power.* London: Longman.

Fairclough, N. (1995) *Critical discourse analysis: The critical study of language.* London: Longman.

Gadamer, H. G. (1975/1994). Truth and method. New York: Seabury.

Gadamer, H. G. (1981). *Reason in the age of science* (F. G. Lawrence, Trans.). Cambridge, MA: MIT Press.

Gadamer, H. G. (1987). The Problem of historical consciousness. In P. Rabinow & W. M. Sullivan (Eds.), *Interpretive social science* (pp. 82–140). Berkeley: University of California Press.

Gee, J. P. (1996). *Social linguistics and literacies: Ideology in discourses* (2nd ed.). London: Taylor and Francis.

Gee, J. P. (1999). *An introduction to discourse analysis: theory and method.* London: Routledge.

Gee, J. P., Hull, G. & Lankshear, C. (1996). *The new work order: Behind the language of the new capitalism.* Boulder, CO: Westview Press.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment.* London: Falmer Press.

Gipps, C. V. (1999). Socio-cultural aspects of assessment. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (Vol. 24, pp. 355–392). Washington, DC: American Educational Research Association.

Greeno, J. (in press). Students with competence, authority and accountability: Affording intellective identities in the classroom. *The College Board.*

Hattie, J., Jaeger, R. M., & Bond L. (1999). Methodological questions in educational testing. In A. Iran-Nejad & P. D. Pearson

(Eds.), *Review of research in education* (Vol. 24, pp. 393–446). Washington, DC: American Educational Research Association.

Heritage, J. (1984). *Garfinkel and ethnomethodology.* Cambridge, UK: Polity.

Hoy, D. C., & McCarthy, T. (1994). *Critical theory.* Oxford: Blackwell.

Johnston, P. H. (1989). *Constructive evaluation of literate activity.* New York: Longman.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342.

Lampert, M. (2001). *Teaching problems and the problems of teaching.* New Haven: Yale.

Lather, P. A. (1991). *Getting smart: feminist research and pedagogy with/in the postmodern.* New York: Routledge.

Lather, P. A. (1999). To be of use: The work of reviewing. *Review of Educational Research, 68*(1), 2–7.

Lather, P. (2001). Validity as an incitement to discourse: Qualitative research and the crisis of legitimation. In Virginia Richardson (Ed.), *Handbook of research on teaching* (4th ed.). Washington, DC: American Educational Research Association.

Lave, J. (1988). *Cognition in practice.* Cambridge, UK: Cambridge University Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge, UK: Cambridge University Press.

Lee, C. D. (2001). Is October Brown Chinese? A cultural modeling activity system for underachieving students. *American Educational Research Journal, 38*(1), 97–141.

Martin, M., & McIntyre, L. C. (1994). Introduction. In M. Martin & L. C. McIntyre (Eds.), *Readings in the philosophy of social science* (pp. xv–xxii). Cambridge, MA: The MIT Press.

Mehan, H. (1993). Beneath the skin and between the ears: A case study in the politics of representation. In S. Chaiklin & J. Lave (Eds.), *Understanding practice: Perspectives on activity and context* (pp. 241–268). Cambridge, UK: Cambridge University Press.

Mehan, H. (1998). The study of social interaction in educational settings: Accomplishments and unresolved issues. *Human Development, 41,* 245–269.

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–24.

Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment.* Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Mislevy, R. J., Almond, R., & Steinberg, L. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229–258.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5–12.

Moss, P. A. (1995). Themes and variations in validity theory. *Educational measurement: Issues and Practice, 14*(2), 5–13.

Moss, P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher, 25*(1), 20–28, 43.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*(2), 5–12.

Moss, P. A., Pullin, D. P., Gee, J. P., & Haertel, E. H. (2002). *The idea of testing: Expanding the foundations of educational measurement.* Unpublished manuscript, University of Michigan.

Moss, P. A., & Rex, L. A. (1997/2001). *Syllabus: qualitative methods in educational research.* Unpublished syllabus, University of Michigan.

Moss, P. A., & Schutz, A. M. (2001). Educational standards, assessment, and the search for consensus. *American Educational Research Journal, 38*(1), 37–70.

Moss, P. A., Sutherland, L. M., Haniford, L., Miller, R., Johnston, D., Geist, P. K., et al. (2003). *Interrogating the generalizability of portfolio assessments of beginning teachers: A qualitative study.* Unpublished manuscript, University of Michigan.

National Research Council. (2001). *Knowing what students know* (J. W. Pellegrino, N. Chudowsky, & R. Glaser, Eds.). Washington, DC: National Academy Press.

Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, NJ: Merrill.

Ormiston, G. L., & Schrift, A. D. (Eds.). (1990). *The hermeneutic tradition: From Ast to Ricoeur.* Albany, NY: SUNY Press.

Patton, M. Q. (2001). *Qualitative research and evaluation methods.* Thousand Oaks, CA: Sage.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). In A. Iran-Nejad and P. D. Pearson (Eds.), *Review of research in education* (Vol. 24, pp. 307–354). Washington, DC: American Educational Research Association.

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life.* Princeton, NJ: Princeton University Press.

Power, B. M., & Hubbard, R. S. (1999). *The art of classroom inquiry.* Portsmouth, NH: Heinemann.

Rabinow, P., & Sullivan, W. M. (1987). *The interpretive turn: A second look.* In P. Rabinow & W. M. Sullivan (Eds.), *Interpretive social science* (pp. 1–32). Berkeley: University of California Press.

Rawls, A. W. (2001). Editor's introduction. In *Ethnomethodology's Program* (H. Garfinkel, Ed.) (pp. 1–64). Lanham, MD: Rowan and Littlefield.

Rex, L. A. (2001). The remaking of a high school reader. *Reading Research Quarterly, 36*(3), 288–314.

Rex, L. A., & McEachen, D. (1999). "If anything is odd, inappropriate, confusing, or boring, it's probably important": The emergence of inclusive academic literacy through English classroom discussion practices. *Research in the Teaching of English, 34*(1), 65–129.

Rogoff, B., & Lave, J. (Eds.) (1984). *Everyday cognition: Its development in social context.* Cambridge, MA: Harvard University Press.

Rogoff, B., Turkanis, C. G., & Bartlett, L. (2001). *Learning together: Children and adults in a school community.* New York: Oxford.

Santa Barbara Classroom Discourse Group. (1992). Constructing literacy in classrooms: Literate action as social accomplishment. In H. H. Marshall (Ed.), *Redefining student learning* (pp. 119–150). Norwood, NJ: Ablex Publishing Corporation.

Schmidt, L. K. (Ed.). (1995). *The specter of relativism: Truth, dialogue, and phronesis in philosophical hermeneutics.* Evanston, IL: Northwestern University Press.

Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed.* New Haven, CT: Yale University Press.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405–450.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5–8.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 1066–1101).

Washington, DC: American Educational Research Association.

Shepard, L. A. (2003). Commentary: Intermediate steps to knowing what students know. *Measurement: Interdisciplinary Research and Perspectives, 1*(2), 171–177.

Taylor, C. S., & Nolen, S. B. (1996, November 11). What does the psychometrician's classroom look like? Reframing assessment concepts in the service of learning. *Educational Policy Analysis Archives, 4,* 17. Retrieved August 7, 2003 from http://epaa.asu.edu/epaa/v4n17/

Thompson, J. B. (1990). The methodology of interpretation. In *Ideology and modern culture* (pp. 272–327). Stanford: Stanford University Press.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity.* Cambridge, UK: Cambridge University Press.

Wertsch, J. V. (1998). *Mind as action.* Oxford: Oxford University Press.

Wertsch, J. V., Del Rio, P., & Alverez, A. (1995). Sociocultural studies: History, action, and mediation. In J. V. Wertsch, P. Del Rio, & A. Alverez (Eds.), *Sociocultural studies of mind* (pp. 1–36). Cambridge, UK: Cambridge University Press.