Validity in Educational Assessment
Author(s): Pamela A. Moss, Brian J. Girard, Laura C. Haniford
Published by: American Educational Research Association
Stable URL: http://www.jstor.org/stable/4129771

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at
http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless
you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you
may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at
http://www.jstor.org/action/showPublisher?publisherCode=aera.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed
date of such transmission.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of
content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms
of scholarship. For more information about JSTOR, please contact support@jstor.org

American Educational Research Association is collaborating with JSTOR to digitize, preserve and extend
access to Review of Research in Education.
Chapter 4

Validity in Educational Assessment

PAMELA A. MOSS
University of Michigan

BRIAN J. GIRARD
University of Michigan

LAURA C. HANIFORD
California State University, Fullerton

Teachers and other education professionals use assessments in different contexts: in classrooms, subject-matter departments, committees, and informal learning groups; in schools; in district, state, and federal education offices; on school boards, commissions, and legislatures; and so on. Professionals working in different contexts have different decisions to make, different sources of evidence, different resources for interpreting the available evidence, and different administrative constraints on their practice. Educational assessment should be able to support these professionals in developing interpretations, decisions, and actions that enhance students’ learning. Validity refers to the soundness of those interpretations, decisions, or actions. A validity theory provides guidance about what it means to say that an interpretation, decision, or action is more or less sound; about the sorts of evidence, reasoning, and criteria by which soundness might be judged; and about how to develop more sound interpretations, decisions, and actions. In this chapter, we review three distinct theoretical discourses—educational measurement, hermeneutics, and sociocultural studies—to support the development of validity theory for the routine use of assessment by professionals working in complex, dynamic, and always partially unique educational environments.1,2

Like any theory, a validity theory can be construed as an intellectual framework or set of conceptual tools that shapes both our understanding and our actions. It illuminates some aspects of social phenomena for consideration and leaves others in the background. As Mislevy (2006) notes, different theories:

are . . . marked by what is attended to and how it is thought about; . . . by the kinds of problems they address, the solutions they can conceive, and the methods by which they proceed; by the “generative principles” (Greeno, 1989) through which experts come to reason in a domain. They are also marked by what is not emphasized, indeed what is ignored. (p. 269)
We represent different validity theories as conceptual tools that enable and constrain what we can understand about the social world and that shape and are shaped by the contexts in which they are used.

A validity theory can also be construed as the representation of an epistemology—a philosophical stance on the nature and justification of knowledge claims—which entails a philosophy of science. The most sustained, explicit effort to develop validity theory about learning assessment has occurred within the field of educational measurement. Validity theory in educational measurement has been, for the most part, grounded in the epistemological understandings of a naturalist or unified conception 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, the primary goals of social science are nomological (law-like) or generalizable explanation or prediction. In contrast, those who take what is often called an "interpretive" approach to social science (Bohman, Hiley, & Shusterman, 1991; Flyvbjerg, 2001; Rabinow & Sullivan, 1987) argue that social phenomena differ from natural phenomena because they are meaningful to the actors involved. Furthermore, meanings are seen as embedded in complex social contexts that shape (enable and constrain) what can be understood in ways that the actors involved may not perceive. From this perspective, a primary aim of social science is to understand what people mean and intend by what they say and do and to locate those understandings within the historical, cultural, institutional, and immediate situational contexts that shape them. We turn to interpretive social science for two additional sets of theoretical resources—hermeneutics and sociocultural studies—to complement, extend, and reframe validity theory in educational measurement, thus situating it within a broader field of epistemological possibilities.

In educational measurement, validity theories have been developed around the use of tests and other standardized forms of assessment. Although validity has focused on interpretations or uses of test scores rather than on the test itself—and scores from a particular test can be interpreted and used in multiple ways—the conceptual architecture is tied to a particular standardized assessment and the circumscribed evidence it provides about student learning. The questions that can be addressed and the interpretations, decisions, and actions that can be supported by the validity evidence provided are similarly circumscribed.

Yet, the problems, issues, and questions that education professionals face and the types of evidence they need to address them are not similarly circumscribed. In describing the "problems of teaching," Lampert (2001) uses the productive metaphor of a camera lens shifting focus and zooming in and out. This allows her to address: "the problems of practice that a teacher needs to work on in a particular moment... [along with] the problems of practice that are addressed in teaching a lesson or a unit or a year" (pp. 2–3). An analogy may be easily drawn to the problems or issues facing professionals in other contexts. Many of the problems are of the "what do I/we do next?" variety, albeit at different levels of scale.
What’s needed is a flexible approach to validity that begins with the questions that are being asked; that can develop, analyze, and integrate multiple types of evidence at different levels of scale; that is dynamic so that questions, available evidence, and interpretations can evolve dialectically as inquirers learn from their inquiry; and that allows attention to the antecedents and anticipated and actual consequences of their interpretations, decisions and actions. [From here on, we will refer to interpretations, decisions, and action, which frequently entail each other in practice, as IDAs, unless we need to refer to them separately.]

We also need to recognize that not all IDAs should or can be subjected to an explicit reflection on or documentation of their validity; in fact, most IDAs fall into this category. Much that might be called assessment is simply a routine part of social interaction in a learning environment. Therefore, we need a set of conceptual tools that can be applied, explicitly when needed, but that can also provide actors with adequate information and rules of thumb to shape their daily practice. Validity theory should also speak to the meta-issue of how learning environments are resourced—with knowledgeable people, material and conceptual tools, norms and routines, and evolving information about learning—to support sound evidence-based IDAs when explicit inquiry is not possible. It should speak to the issues of how assessment supports the professionals’ learning to support students’ learning and one another’s learning (Moss & Greeno, in press).

Finally, we must recognize that assessment practices do far more than provide information, they also shape people’s understanding about what is important to learn, what learning is, and who learners are (Engeström, 2001; Lave, 1993). Thus, any validity theory needs to consider how assessment functions as part of—shaping and shaped by—the local learning environment and its learners. We draw on hermeneutics to suggest practices for warranting IDAs that integrate multiple types of evidence, that evolve as new evidence is brought to bear, and that illuminate the social forces that shape them. We draw on sociocultural studies to suggest types of evidence that might be considered in developing an IDA and to provide resources for analyzing complex and dynamic learning environments in which assessment practices function.

In the following sections, we review validity theories in educational measurement, paying particular attention to how the needs of and expectations for the local user of standardized assessments are considered in this literature. Next, we expand the conception of assessment, pointing to studies of assessment in use and to the implications for assessment in evolving conceptions of learning. Then, we turn to theoretical resources for validity theory that can be located within an interpretive approach to social science and that support the development of situated interpretations. We turn to hermeneutics and its potential for supporting/inquiring into the validity of IDAs in response to a range of educational issues, problems, or questions. We also turn to sociocultural studies and their potential for analyzing complex and dynamic learning environments and the interactions between them (such as when evidence and people cross boundaries between levels of the educational system). We use extended examples as transitions between sections to demonstrate the need for alternative theoretical
resources and to illustrate the theoretical resources that precede or follow. The examples illuminate some of the issues involved in developing validity theory to support assessment (1) as it might be used within a focal learning environment (e.g., a classroom) by those faced with the decision of “what do we do next?”; (2) as it might be used outside the focal learning environment by others (e.g., administrators, other teachers, and policy makers) who are responsible for monitoring, supporting, or evaluating the focal learning environment; and (3) as it might be used by those who are seeking to learn about the design and support of learning environments and the assessment practices enacted within them. We close with an analytic overview of the sets of theories on which we have drawn and suggest next steps for research and practice.6

VALIDITY IN EDUCATIONAL MEASUREMENT

Validity theories in educational measurement have been developed, primarily, to evaluate intended interpretations and uses of scores on tests or other standardized forms of assessment. Scores summarize patterns or consistencies in performance.7 The scores and their aggregates are interpreted and used to inform decisions and actions. Standardization allows a common validity inquiry to be developed and used across individuals and contexts for which the assessment was developed. Validity theories in educational measurement offer principles, practices, and types of evidence through which interpretations associated with scores (and their implications for decision and action) should be evaluated. Much of the evidence and rationale supporting the validity of the intended interpretation is developed before the first operational use of a test (although validity inquiry often continues) and assumptions are made about the meaningfulness and appropriateness of this validity evidence in similar contexts with similar individuals. Educational measurement validity theorists also routinely caution that assessment developers and users have the obligation to identify those cases where the common interpretation and validity inquiry do not hold.

In this section, we first provide a brief history of the validity concept in educational measurement as traced in seminal documents to set what follows in context and to provide background for readers new to the field.8 Then, we present descriptions of two more current approaches to validity theory/practice: one we call validity theory as scientific inquiry (e.g., Cronbach, 1988, 1989; Messick, 1989) and one we call validity theory as practical argument (e.g., Kane, 2006; Mislevy, Almond, & Steinberg, 2003)—labels taken from language used by the theorists. Although these approaches can be seen as largely complementary from an epistemological perspective, with the latter evolving from the former, viewed from the perspective of conceptual tools there are differences that potentially hold substantial implications for shaping understanding and practice. Finally, we consider the relevance of validity theories in educational measurement for teachers, administrators, and other education professionals.

Developments in validity theory in educational measurement can be traced to successive editions of two seminal publications: (1) the Standards for Educational
and Psychological Testing (Testing Standards) (as it is now called), jointly sponsored by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (Testing Standards, for short), and (2) the “Validity Chapter” in Educational Measurement, sponsored by the National Council on Measurement in Education and the American Council on Education (Validity Chapter, for short).9

Each edition of the Testing Standards is drafted by a committee of measurement scholars jointly appointed by the sponsoring organizations. The Testing Standards are intended to provide specific guidelines to test developers and users about “criteria for the evaluation of tests, testing practices, and the effects of test use” (1999, p. 2) and to represent a current “consensus” about best practice within the measurement field.10 The process of developing an edition generally takes multiple years as developing drafts are submitted to the sponsoring organizations and the field for review and comment. Five editions of the standards have been published—in 1954/1955, 1966, 1974, 1985, and 1999—under slightly different names.11

The Validity Chapters have served a somewhat different purpose. These chapters provide historical overviews of validity theories, typically situating them within the philosophies of social science from which they are derived, and provide elaborated statements of the authors’ views about how validity should be represented and practiced. The authors are typically selected for their sustained and seminal contributions to validity theory by an editorial board. Although multiple drafts of the chapters are peer reviewed by selected scholars in the field, the chapters are expected to represent the individual authors’ perspectives on validity. There have been four Validity Chapters written since 1950, by Edward E. Cureton (1951), Lee J. Cronbach (1971), Samuel Messick (1989), and Michael J. Kane (2006).


In the first Validity Chapter, Cureton (1951) characterized validity as indicating “how well the test serves the purpose for which it is used” (p. 621). He operationalized validity in terms of the relationship between test scores and “criterion” scores, which reflected performance “on the actual task” (p. 622) the test is used to measure. He argued that “[s]ince a test can be used as an indicator or predictor of performance in any number of situations, it can have as many validities . . . as the number of different criteria with which it is correlated” (p. 625). The chapter is rich with largely hypothetical examples of how one might develop criterion measures for concepts as varied as generosity, kindness, typing, proficiency in flying, and success in medical practice or engineering, along with various educational aims, such as artistic appreciation, reasoning ability, and effective use of the English language.12

By the time the first edition of the Testing Standards was published in 1954, the concept of validity had been conceptualized in terms of different “aims” of testing, each associated with a different “type” of validity investigation (of which Cureton’s criterion-related validity was only one). This conception of validity persisted in similar form through the first three editions of the Testing Standards (1954/1955,13
1966, and 1974). The authors of the 1966 Testing Standards characterized three aims: (1) determining "how an individual performs at present in a universe of situations the test situation is claimed to represent" (e.g., an achievement test), where "content validity" demonstrates "how well the test samples the class situations or subject matter about which conclusions are to be drawn"; (2) "forecasting an individual future standing or estimating an individual's present standing on some variable of particular significance" (as when a test of academic aptitude is used to predict grades) where "criterion-related validity" compares test scores with "one or more external variables considered to provide a direct measure of the characteristic or behavior in question"; and (3) inferring "the degree to which the individual possess some hypothetical trait or quality (construct) . . . that cannot be observed directly" (such as intelligence or creativity) (p. 13), where "construct validity" involved "a combination of logical and empirical attack" to investigate what qualities a test measures (pp. 12–13).

The concept of construct validity, elaborated by Cronbach and Meehl (1955), was viewed at the time as an "indirect" method of validation to be used when no criterion variable or content domain could indicate the degree to which a test measured what it was intended to measure, although this conception of validity soon became central. As fleshed out in the 1966 Standards:

Essentially, studies of construct validity check on the theory underlying the test. The procedure involves three steps. First, the investigator inquires (sic): From this theory, what hypotheses may we make regarding the behavior of persons with high and low scores? Second, he gathers data to test these hypotheses. Third, in light of the evidence, he makes an inference as to whether the theory is adequate to explain the data collected. If the theory fails to account for the data, he should revise the test interpretation, reformulate the theory, or reject the theory altogether (American Psychological Association [APA], 1966, pp. 12–13).

In his 1971 Validity Chapter, Cronbach gave "construct validity" far more centrality in his general conception of validity than had the Standards. Although he maintained the relevance of the now-familiar three types of validity inquiry (content, construct, and criterion-related) as aspects of a validity inquiry, he likened validity research to the evaluation of a scientific theory as characterized in "construct validity," and he argued that most educational tests entailed constructs: "whenever one classifies situations, persons, or responses, he uses constructs" (p. 462).

The 1985 Testing Standards moved in this direction, articulating more prominently a unified conception of validity that draws on multiple types of evidence to evaluate the inferences and uses of test scores. The authors renamed the traditional three categories to emphasize their role as types of evidence (construct-, content-, and criterion-related evidence) rather than types of validity: "An ideal validation includes several types of evidence, which spans all three of the traditional categories" (p. 9).

The third Validity Chapter (Messick, 1989) represented the completion of the move to a unified conception of validity as scientific inquiry into score meaning as articulated in the description of construct validity. It also drew on Messick's earlier work to high-
light the importance of understanding the value implications of score meaning and the consequences of test use as part of this research effort. Thus one can trace across the seminal documents an evolution in the consensus understanding of validity within the measurement field and the implications for what sort of evidence should be made available to enable sound interpretations and uses. Messick characterized this as a: “shift in emphasis from numerous specific criterion validities to a small number of validity types and finally to a unitary validity conception” (p. 18), which privileged the explanation of score meaning as the fundamental focus of validity.

Validity Theory as Scientific Inquiry

In the third edition of the Validity Chapter, Messick (1989) argued that validation “embraces all of the experimental, statistical, and philosophical means by which hypotheses and scientific theories are evaluated” (p. 14). He defined 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” (p. 14). Validation entails ascertaining: “the degree to which multiple lines of evidence are consonant with the inference, while establishing that alternative inferences are less well supported” (p. 13). It also entails: “appraisals of the relevance and utility of test scores for particular applied purposes and of the social consequences of using the scores for applied decision making” (p. 13).

Drawing on accumulated scholarship in the field, Messick articulated general guiding principles and categories of evidence that he argued should be considered in validity inquiry: “Convergent” evidence indicates test scores are related to other measures of the same construct and to other variables they should relate to as predicted by the conceptual framework; “discriminant” evidence indicates that test scores are not unduly related to measures of other constructs. For instance, within some conceptual frameworks, scores on a multiple-choice reading comprehension test might be expected to relate more closely (convergent evidence) to other measures of reading comprehension, perhaps using other more extended texts or response formats (e.g., retelling or thinking aloud). Conversely, test scores might be expected to relate less closely (discriminant evidence) to measures of the specific subject matter knowledge reflected in the reading passages on the test, indicating it is not primarily a measure of topic knowledge.

Construct validation is most efficiently guided by the testing of “plausible rival hypotheses,” which suggest credible alternative explanations or meanings for the test score that are challenged and refuted by the evidence collected. Prominent rival hypotheses or threats to construct validity include “construct underrepresentation” and “construct-irrelevant variance.” “Construct underrepresentation” refers to a test that is too narrow in that it fails to capture important aspects of the construct. “Construct-irrelevant variance” refers to a test that is too broad in that it requires capabilities irrelevant or extraneous to the proposed construct. Continuing with the previous
example, a potential rival hypothesis to the claim that a test measures reading comprehension is that it depends unduly on knowledge of specialized vocabulary, thus reflecting "construct irrelevant variance." Another rival hypothesis might be that the test "underrepresents" the construct by focusing on literal recall to the exclusion of interpretation or by using only short reading passages that are not typical of those students encounter in or out of school. These rival hypotheses suggest studies that might be conducted as part of the construct validation effort.

Messick (1989) also offered readers a list of sources of evidence that he argued should be gathered in the process of validity research, acknowledging that different types of inferences may require different emphases in evidence:

We can [a] look at the content of the test in relation to the domain [about which inferences are to be drawn] . . . , [b] probe the ways in which individuals respond to items or tasks . . . , [c] examine relationships among responses to the tasks, items, or parts of the test, that is the internal structure of test responses . . . , [d] survey relationships of the test scores with other measures and background variables, that is, the test's external structure . . . , [e] investigate differences in these test processes and structures over time, across groups and settings, and in response to experimental interventions, [and] [f] trace the social consequences of interpreting and using the test scores in particular ways, scrutinizing not only the intended outcomes but also unintended side effects. (p. 16)

To evaluate the potential consequences of a test use, Messick suggested pitting the proposed use against alternative assessment techniques and alternative means of serving the same purpose, including the generalized alternative of not assessing at all. This juxtaposition of alternatives exposes: "the value assumptions of a construct theory and its more subtle links to ideology" (p. 62). Messick also suggested contrasting "multiple value perspectives," citing the example of different value perspectives underlying various selection systems (e.g., according to abilities, efforts, accomplishments, and needs). These contrasts provide opportunities for well-informed debate.

As Messick (1989) noted, construct validation rarely results in a summative decision about whether a given interpretation is justified. More typically, the outcomes of a given study or line of research result in the modification of the test, the construct, the conceptual framework surrounding the construct, or all three. Thus, construct validity is as much an aspect of test development as it is of test evaluation. Almost any information gathered in the process of developing and using an assessment is relevant to construct validity when it is evaluated against the theoretical rationale underlying the proposed interpretation. Messick also acknowledged the need to balance the never-ending aspect of this approach to validity inquiry alongside the need to make "the most responsible case to guide . . . current use of the test" (p. 13).

In her 1993 Review of Research in Education (RRE) review "Evaluating Testing Validity," Shepard raised two interrelated concerns about Messick's conception of validity as scientific inquiry. First, she worried about the implication that research into score meaning precedes consideration of test use, misdirecting the conceptualization of theoretical frameworks intended to guide validity evaluations. Second, she worried that this approach, taken together with the conceptualization of valid-
ity as a never-ending process, may give “the sense that the task is insurmountable” and allow “practitioners to think that a little bit of evidence of whatever type will suffice” (p. 429). She argued, “measurement specialists need a more straightforward means to prioritize validity questions. If a test is proposed for a particular use, a specific network of interrelations should be drawn focused on the proposed use” (p. 429). She proposed that validity evaluations be organized in response to the question “What does the testing practice claim to do?” She raised additional questions: “What are the arguments for and against the intended aims of the test? And, what does the test do in the system other than what it claims, for good or bad?” (p. 429). A major contribution of her chapter is the construction of extended validity cases to illustrate her theoretical perspective. She pointed to Cronbach’s (1988, 1989) conception of validity as evaluation argument, with its ties to program evaluation, and to a seminal article by Michael Kane (1992) that fleshed out this perspective of validity as interpretive or practical argument. Although Kane’s approach is consistent with Messick’s, it draws our attention more explicitly to the chain of inferences needed to move from the specific performances on a particular test to the “target domain” about which we want to draw conclusions and the decisions those conclusions inform.

Validity as Practical Argument

In his 2006 Validity Chapter, Kane criticized validity theory as being “quite abstract” and called for a more pragmatic approach to validation. He defined validation “as the process of evaluating the plausibility of proposed interpretations and uses” and validity “as the extent to which the evidence supports or refutes the proposed interpretations and uses” (p. 17). The same test score “may have several legitimate interpretations and may be used to make different kinds of decisions” (p. 29), each of which requires validation.

He conceptualized validation as employing two kinds of argument: (1) An interpretive argument lays out “the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances” (p. 23). It provides the framework for the validity argument. (2) The validity argument draws on logical and empirical evidence to provide an evaluation of the overall coherence of the interpretive argument and the plausibility of each of its assumptions. Importantly, he noted that specifying an interpretive argument helps protect against inappropriate interpretations and uses by making gaps in the evidence harder to ignore. He noted, for instance:

The target domains of most interest in education are not restricted to test items or test-like tasks, although they may include this kind of formal performance as a subset. A person’s level of literacy in a language depends on his or her ability to perform a variety of tasks in a variety of contexts, ranging from the casual reading of a magazine to the careful study of a textbook or technical manual. These performances can occur in a variety of locations and social situations . . . The match between the target domain and the measure of a trait is a central issue in developing a trait measure and in validating a trait interpretation. (2006, p. 31)
Kane (1992) described categories of inferences that appear regularly in interpretive arguments. The first three (scoring, generalization, and extrapolation) are involved in almost all interpretive arguments.18

- The scoring inference involves the assignment of a score to each person’s performance. This might entail scoring a student’s responses to individual problems and then combining them into a total score.
- The generalization inference extends the interpretation from the performances actually observed to the “universe of generalization” composed of performances on similar test-like tasks under similar circumstances.
- The extrapolation inference extends the interpretation again from the universe of generalization to the much broader “target domain” (trait) about which the test user wants to draw inferences.
- An implication inference extends the interpretation to include any claims or suggestions that might be associated with verbal descriptions of the test score.
- A decision inference links the test scores to any decisions or actions and the potential (intended and unintended) consequences associated with them.
- A theory-based inference extends the interpretations to the underlying mechanisms or relationships that account for the observed phenomena.
- Technical inferences involve the appropriateness of assumptions regarding technical issues, such as equating forms, scaling, and the fit of statistical models.

The details of the interpretive argument depend on “the specific interpretation being proposed, the population to which the interpretation is applied, the specific data collection procedures being used, and the context in which measurement occurs” (1992, p. 529). Interpretive arguments can involve a number of inferences, some of which can be taken for granted “unless special circumstances suggest otherwise” (e.g., that students can read the question) and some of which will need to be evaluated (e.g., that an achievement test covers a content domain, which, he notes, is almost always questionable) (2006, p. 23).

The validity argument provides an evaluation of the interpretive argument. It would: “begin with a review of the argument as a whole to determine if it makes sense” (2006, p. 25). Then it would proceed through an evaluation of each of the inferences, each one involving multiple types of evidence, “including expert judgment, empirical studies, the results of previous research, and value judgments” (2006, p. 25). He divided the process of validation into two stages, one that occurs during the development of a test (the development stage), where any weaknesses identified can lead to modification of the test or the interpretive argument, and one that occurs after a test is developed (the appraisal stage), which is intended to be far more critical. Although low-stakes tests may only entail evidence gathered during the development stage; high-stakes tests appropriately require a more extensive evaluation of the fully developed test in use.
Kane argued that interpretive arguments tend to rely on informal or “presumptive” reasoning (citing Toulmin, 1958): their claims “are always somewhat tentative, and often include explicit indications of their uncertainty” (2006, p. 27) (e.g., standard errors). If the interpretive argument survives all reasonable challenges, it can be accepted as provisionally valid with the understanding that new evidence may later undermine its credibility. He noted further that: “the generic form of the interpretive argument represents the proposed interpretations and uses of test scores. It is applied every time test results are used to draw conclusions or make decisions” (2006, p. 25), such that it does not have to be developed anew for each person. However, “even if the interpretive argument works well in most cases, it may fail in situations in which one or more of its assumptions fails to hold” (p. 25).19 Thus, it falls on local interpreters/users of test-based information to determine the validity of the intended interpretation and use in the local circumstances.

Validity Inquiry and Local Users/Interpreters of Externally Developed Tests

The approaches to validity inquiry/argument described all focus on the intended interpretations and uses of test scores—the interpretations and uses that are “presumptively” valid across individuals and contexts. As Cronbach notes, however:

In the end, the responsibility for valid use of a test rests on the person who interprets it. The published research merely provides the interpreter with some facts and concepts. He has to combine these with his other knowledge about the persons he tests and the assignments or adjustment problems that confront them, to decide what interpretations are warranted. (1971, p. 445)

Those who develop and evaluate assessments “meet [their] responsibility through activities that clarify for a relevant community what a measurement means and the limitations of each interpretation” (Cronbach, 1988, p. 3). The Testing Standards have spoken directly to the issue of what test developers should provide in Technical and Users manuals to enable sound judgments by users. Although generally acknowledging the never-ending nature of validity research, they have spoken about what should be available before the first operational use of a test. The 1999 Testing Standards, drafted just after the publication of Kane’s 1992 article and Messick’s Validity Chapter, walked a middle ground between validity as scientific inquiry and validity as practical argument with the concept of “scientific argument”: “Validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use” (p. 9).20 Similar to Messick, the authors listed five sources of evidence that might inform validity arguments; these were evidence based on “test content,” “response processes,” “internal structure,” “relations to other variables,” and “consequences of testing” (pp. 11–16). The authors of the Testing Standards acknowledged the role of “professional judgment” in deciding what constellation of validity evidence is necessary for a particular interpretation and use, and the document is, consequently, less prescriptive than some would prefer (e.g., Messick, 1989) about what should be included in a scientifically sound validity argument.21
Numerous measurement theorists have noted the disjunction between the scholarly literature and the practice of validity research, at least as reflected in technical and interpretive manuals made available to users. Cronbach (1989) criticized test manuals for providing readers with a “do-it-yourself kit of disjoint facts” (p. 156)—where “validation consists not so much in questioning the proposed interpretation as in accumulating results consistent with it” (p. 152)—rather than incisive challenges to score meaning integrated into an overall validity argument. Messick (1989) and Shepard (1993) echoed similar concerns. Shepard (2006) pointed to the unfortunate tendency of test developers to match test items to standards without illuminating the ways in which tests do and do not cover the depth and breadth of learning implied in the standards, an issue that Kane’s (2006) conception of validity directly confronts.

Since the 1974 edition, the Standards have also included one or more additional chapters aimed at uses of tests, including chapters on educational uses in 1985 and 1999. These chapters offer relevant elaborations on the general concepts in earlier chapters, sometimes specifying additional obligations for users. They also provide some limited guidelines about appropriate practice for users of assessment.

With respect to the expectation that users will interpret test scores in light of additional evidence about the individual case or context, the approaches to validity described have less to offer about how to warrant local interpretations. Advice provided in the 1999 Testing Standards is somewhat ad hoc (and circular):

When interpreting and using scores about individuals or groups of students, consideration of relevant collateral information can enhance the validity of the interpretations, by providing corroborating evidence or evidence that helps explain student performance. Test results can be influenced by multiple factors including institutional and individual factors such as the quality of education provided, students’ exposure to education (e.g., through regular school attendance), and students’ motivation to perform well on the test. (p. 141)

The only “standard”—or explicit obligation—offered with respect to the use of collateral information in the education chapter involves the situation when consequential decisions are made: “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). Examples cited include school records, classroom observations, and parent input. Yet, approaches to validity described so far have little to offer about how to combine such situated evidence into a well-warranted interpretation or decision.

Validity Inquiry and the Local Assessment Developer

The approaches to validity research we described have also been assumed to be relevant to the local assessment developer, including classroom teachers, albeit with somewhat less relevance. The authors of the 1999 Testing Standards assert:
Although the Standards applies 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 . . . instructor-made tests that are used to evaluate student performance in education and training (p. 3).

The authors of the Standards assert further that the standards may have some relevance for even less formal and less standardized "aids to day-to-day evaluative decisions" (p. 3), although they acknowledge that "it would be overreaching to expect that the standards of educational and psychological testing be followed by those making such decisions" (p. 4).

In an extensive historical review of measurement textbooks intended for teachers from the 1940s through the 1990s and of the research literature in measurement on the preparation of teachers, Shepard (2006) concluded: "Measurement theorists, responsible for 'Tests and Measurements' courses for teachers, believed that teachers should be taught how to emulate the construction of standardized achievement tests as well as how to use a variety of standardized measures" (p. 625). She noted that there was little guidance "about how teachers were to make sense of assessment data so as to redesign instruction" (p. 625).23 As we demonstrate in the next sections, the issues teachers, administrators, and other local users of assessment face are considerably broader than what that can be easily addressed within the conventional conceptions of validity and standardized forms of assessment.

Contributions of Educational Measurement
Validity Theory to Educational Assessment

Our primary focus is on the conceptualization of validity theory for the routine use of assessment by professionals working within their own educational contexts. Validity theory in educational measurement contributes in numerous ways to this goal. Consistent with its heritage in a naturalist or unified approach to social science, validity theory in educational measurement supports the development and evaluation of interpretations based on standardized forms of assessment that are intended to be generalizable—meaningful and useful—across relevant individuals and contexts. Validity research is conducted, in part, to ascertain the extent to which such generalizations may be warranted. However, validity theory in educational measurement also acknowledges the presumptive nature of the intended interpretations, the importance of illuminating those cases where the intended interpretation is not viable, and the role of the local interpreter in deciding what interpretation is warranted in a particular case.24

For the developer or evaluator of assessments intended for large scale use, validity theory in educational measurement suggests what evidence and rationale should be made available to support local interpretations and uses (although several theorists have raised concerns about the disjunction between theory and practice). With validity as scientific inquiry, we see a primary emphasis on explaining what scores mean by situating them in a larger theoretical framework. Validity questions about the relevance of the test for serving particular purposes in particular types of contexts are seen
to build on this base. With validity as practical argument, we see a pragmatic move to
focus validity inquiry directly on the purposes for which a test is used (and, we would
argue, the meaning implied in that use). The conceptual architecture provided by
Kane (1992, 2006) illuminates the inferential steps an interpreter must take in mov-
ing from particular observations permitted by the test through the domain of test-like
items to the target domain about which one wants to draw inferences (advice which
echoes Cureton, 1951).25 Although consistent with Messick’s approach to validity, it
draws our attention to a particular aspect of validity work and a kind of evidence that
seems particularly crucial in drawing sound interpretations based on test scores.

Although validity theory in educational measurement has much less to say about
the how one might support situated interpretations of test scores, it does point to
some general principles that we will see reflected, as well, in interpretive approaches
to social science, including the value of challenges to developing interpretations, mul-
tiple sources of evidence, attention to consequences, and, increasingly, extended cases
of practice that both illustrate, and, as Mislevy noted (personal communication), con-
tribute to validity theory.

EXPANDING CONCEPTIONS OF ASSESSMENT

Evolving conceptions of learning and studies of how evidence of learning is actu-
ally used are pushing our conceptions of what assessment is and what validity theory
needs to accomplish. In this section, we illuminate a range of purposes and practices
regarding the use of evidence of student learning that can or have been considered
“assessment” to substantiate the need for an expanded conception of validity. To illus-
trate what validity theory needs to accomplish, we point to studies of assessment in use
in schools, districts, and classrooms; then we consider, briefly, conceptions of learning
and assessment from cognitive and sociocultural theorists; we close with an extended
example of how evidence is used by one teacher to inform her practice (Lampert,
2001). These practices and perspectives point to the importance of being able to rou-
tinely address the validity of IDAs that draw on multiple types of evidence from the
local environment and the importance of understanding how assessment functions
in the social systems of which it is a part.

Use of Assessment in Context

A small but growing body of policy research on how evidence, including stan-
dardized assessment, is used to inform educational practice suggests a range of pur-
poses and practices (e.g., Diamond & Spillane, 2004; Honig & Coburn, 2005; Kerr,
Marsh, Ikemoto, Darilek, & Barney, 2006; Massel & Goertz, 2002; Young, 2006).
Questions to which assessment information is put include those about professional
development planning, curriculum and text-book adoption, targeting topics in need
of improved instruction, developing more specific instructional plans for struggling
students (as well as identifying them and monitoring their progress), setting new
instructional goals and targets, and as a way to communicate with the community in
general and parents in particular (Diamond & Spillane, 2004; Heritage & Yeagley, 2005; Kerr et al., 2006; Massel & Goertz, 2002; Supovitz & Klein, 2003). Serving these purposes frequently requires multiple types of evidence.

Ikemoto and Marsh (in press), reviewing evidence from surveys and interviews of teachers and administrators in 10 districts throughout four states, characterize both data types and data-driven analyses along a continuum from basic to complex (and they are careful to note that complex forms are not always better than basic forms). Data can vary along this axis, depending on several characteristics, including the time frame it represents (single point vs. trends), the amount and sources of data (single score or source vs. multiple types and sources), and whether it is readily available or collected for inquiry purposes. Similarly, data analysis and decision-making practices range from “basic” (e.g., decisions based directly on reported test scores, with minimal additional evidence or analyses) to somewhat more complex analyses of multiple types of evidence to what they call “inquiry focused” analysis, which involves the development of questions and evidence to address them. Other variations along this axis include whether the process is individual or collective, one-time or iterative, based on assumptions or empirical evidence, and using basic or expert knowledge.

Coburn and Talbert (2006), drawing on surveys, interviews, and observations within a single district, uncovered a range of beliefs about what constitutes high-quality evidence (from the range of assessments employed by the district) and how it should be used. For example, some teachers and administrators privileged the alignment of the test with the standards when considering the validity of the tests, whereas others believed that only instruments that allowed them to “see” the students’ cognitive processes were warranted as valid. Some used summative assessment in the service of accountability practices (both internal and external), and others used formative assessments to aid in the placement of students in educational programs and to inform instructional design. As Coburn and Talbert explained, “Individuals’ conceptions of valid evidence, of evidence use, and of high-quality research differed in part according to their location in the local education hierarchy. This pattern appears to stem from the fact that different positions carry different kinds of responsibilities that shape individuals’ conceptions” (p. 482). These differences in practices of data use and conceptions of validity suggest that interpretations and uses are always, ultimately, situated in local contexts that influence their meaning and the consequences of their use.

Shepard (2006) provided a brief review of a limited literature within educational measurement on assessment use at the classroom level:

Several studies, focused initially on how standardized tests were used in classrooms, revealed the much greater importance for teachers’ day-to-day decision-making of teacher-made tests, curriculum-embedded tests, and informal interactions and observations (Dorr-Bremme, 1983; Salmon-Cox, 1981; Yeh, Herman, & Rudner, 1981). From interview data, Dorr-Bremme (1983) concluded that teachers act as practical reasons and as clinicians, orienting their assessment activities to the practical tasks they have to accomplish in everyday routines, such as ‘deciding what to teach and how to teach it to students of different achievement levels; keeping track of how students are progressing and how they (the teachers) can appropriately adjust their teaching; and evaluating and grading students on their performance’ (p. 3) [quotes in original]. (Shepard, 2006, pp. 625–626)

Teachers . . . must detect the needs of individual students and various groups of students, not just at one moment, but continuously, and not for just a few students, but for as many as 150. They use classroom assessments to clarify achievement expectations for their students. Teachers must compile evidence to inform such decisions as the assignment of grades and the identification (with justification) of students with special needs. They use assessment as a behavior control system or motivator, and to evaluate the effectiveness of their instructional interventions, among other uses. (Stiggins, 2001b, p. 9)

Stiggins noted that teachers used various formal and informal assessment methods, including those that relied on personal interaction with students, and many different means of communicating with students about their achievement that were more or less specific (e.g., descriptive or judgmental, narratives, or grades). He noted several variables that related to the “assessment environments” teachers created, including their knowledge of the subject matter, needs for structure and control and willingness to take risks, beliefs about their own role as teachers, and perceptions of their students’ willingness and ability to learn.26 Again we see the importance of being able to address the validity of IDAs that draw on multiple types of evidence, the complex ways in which assessment practices (and hence validity issues) can differ, and the importance of understanding how any assessment is situated in the local context.

Evolving Conceptions of Learning and the Implications for Assessment

Evolving conceptions of learning, especially those based in cognitive and sociocultural studies, have substantial implications for the practice of educational assessment. Although these two research discourses privilege different conceptions of learning, they also share much in common in their visions about the role assessment can and should play in a learning environment.

For instance, in “Knowing What Students Know,” the National Research Council (NRC) (2001) called for an approach to assessment that would “be largely determined by a model of cognition and learning that describes how people represent knowledge and develop competence in the domain” (p. 178). They called for assessments that address learners’ knowledge structures, metacognition, problem-solving strategies, progress along a domain specific developmental continuum, transfer of knowledge and skills, and problems in multiple contexts, and communicative practices in the domain. Assessments, they argued, make students’ thinking visible to teachers and themselves, allow timely and informative feedback to enable learning, and show students’ progress along a domain-specific developmental continuum (pp. 3–5).27

Sociocultural studies of learning construe learning in the interaction between learners and their environments, including all the conceptual tools (such as languages), physical tools (such as computers or dictionaries), and other people, all of which serve as resources and constraints on learning (Gee, forthcoming).28 Whereas
the emphasis in cognitive studies of learning is on “the way knowledge is represented, organized, and processed in the mind” (NRC, 2001, p. 3), the emphasis in sociocultural studies of learning is on participation in a practice where the knowledge is meaningful and useful (Greeno, 2002; Greeno and Gresalfi, forthcoming).

Learning also entails becoming a person for whom—and a member of a community in which—such knowledge is meaningful and useful. Thus, as Lave and Wenger (1991) put it, “learning involves the construction of identities” (p. 53). One can take on different identities in different communities that offer different resources for learning. Focusing on classroom identities, Greeno (2002) articulated an important distinction between conceptual and interpersonal relations and the different positions afforded individuals. Interpersonal relations “involve attention to the ways individuals are entitled, expected, and obligated to act toward each other” (p. 3). Conceptual relations involve “attention to the ways that individuals, the class, and groups within the class are entitled, expected, and obligated to act in and toward the subject matter content of the class” (Greeno, 2002, pp. 3–4). Again, this suggests that evidence of learning entails evidence of the interaction between learners and their environment.

Caroline Gipps, in her 1999 RRE chapter on sociocultural perspectives on assessment, highlighted the implications sociocultural theories of learning have for the practice of assessment: “. . . the requirement is to assess process as well as product; the conception must be dynamic rather than static; and attention must be paid to the social and cultural context of both learning and assessment” (p. 375). This includes the local context, as well as the larger institutional and social contexts in which it is embedded which shape the IDAs of local actors. These are themes which, in somewhat different language, appear throughout multiple articulations of sociocultural and sociological approaches to assessment (Broadfoot, 1996; Delandshere, 2002; Filer, 2000; Gipps, 2002; Hickey & Zuiker, 2003; Lee, forthcoming; Mehan, forthcoming; Moss & Greeno, forthcoming; Moss, Pullin, Haertel, & Gee, 2005).29 Put more simply, the goal is to assess what, why, and how students are learning (Engeström, 2001; National Academy of Education [NAE], 2005). This entails understanding the relationship between learners and their learning environment and the larger social systems within which learning is occurring.

The importance of “formative assessment,” or “assessment for learning”30 is prominently represented in both cognitive and sociocultural approaches to learning. Shepard (2006), drawing on both discourses, defined formative assessment as “assessment carried out during the instructional process for the purpose of improving teaching or learning” (p. 627). She included the following as examples of assessment tools: oral questioning of students, observation, written work products, oral presentations, interviews, projects, portfolios, tests, and quizzes. She noted that the strategies could be either formal, where students know they are being assessed, or informal, where assessment is done in the ongoing context of instruction (Shepard, 2006; Shepard et al., 2005). She argued that an equally important role for classroom assessment is evaluation of teaching, where teachers consider “which . . . practices are working and
which are not, and what new strategies are needed” (2006, p. 634). The evaluation of teaching, she noted, asks about the adequacy of students’ opportunities to learn. While acknowledging the institutional role of summative assessment for documenting achievement, she described some of the ways it can pose threats to the type of learning that formative assessment supports. She offered suggestions for using summative assessment in a way that is coherent and that supports the developing competence of students toward valued learning goals. These arguments point us to the importance of integrating multiple types of evidence about interactions among learners, teachers, and resources, including evidence of how assessment is shaping learning.

Delandshere (2002) and Moss and Greeno (forthcoming; Moss, 2003) suggested that a sociocultural approach to assessment challenges our conception of assessment as a distinct component of a learning environment.

We are moving here from an educational practice of assessment where we have defined a priori what we are looking for, to an educational practice where we are participating in activities in which we formulate representations to better understand and transform the world around us. If our purpose is to understand and support learning and knowing and to make inferences about these phenomena, then it seems that the idea of inquiry—open, critical, and dialogic—rather than of assessment (as currently understood) would be more helpful. (Delandshere, 2002, p. 1475)

Further Expanding Our Conception of Assessment

Although traditional distinctions between formative and summative assessment, or between assessment and learning activities, can be useful, it is important to remember that they conglomerate features and reify the resulting practices in arbitrary ways. To illustrate the complexities of how evidence is used in classroom learning environments, we turn to the case of Magdalene Lampert’s teaching practice. We draw on Lampert’s (2001) book, Teaching Problems and the Problems of Teaching, in which she analyzes her practice throughout 1 year of teaching mathematics to a diverse group of fifth graders. Given the goals of this chapter, we read Lampert’s work with the question of how she uses evidence of students’ learning to inform her practice.31

Lampert’s conception of learning mathematics involves a particular understanding of the content—what it means to do mathematics—and of the identities and positions this entails. Lampert’s curriculum was organized around mathematical problems, both practical ones—“like figuring out prices and schedules,” and intellectual ones, “like identifying the conditions that make it necessary to conclude that all of the numbers that are divisible by twenty-one are also divisible by seven” (p. 6). “In the course of working on problems,” Lampert reports, “[students] investigated different solution strategies,” “represented relationships graphically and symbolically,” and “disagreed and defended their approaches and clarified their assumptions” (p. 5). A typical day in Lampert’s classroom began with one or more problems on the board, which students copied into their notebooks, worked on alone or in small groups, and then presented and discussed their work as a whole class. “Learning in my class,” she states, “was a matter of becoming convinced that your strategy and your answer are mathematically legitimate” (p. 6).
As Lampert notes, this conception of learning entails the development of an academic identity, of becoming someone "more inclined to study, to initiate the investigation of ideas, and to be identified as someone who will and can do what needs to be to done to learn in school" (p. 268). Her teaching thus involves "influencing students to be the kinds of persons who are academic resources for themselves and for one another" (p. 266). She notes how being able "to make a mistake [in front of ones peers], admit one has made it, and correct it" (p. 266) is an essential part of an academic character.

Just like her curriculum, Lampert's representation of her teaching—and the evidence it entails—is similarly organized around problems. As noted in the introduction, Lampert uses the productive metaphor of a camera lens shifting focus and zooming in and out to represent the questions that shape her practice. She also notes that multiple problems—sometimes with conflicting solutions—must be solved at once.

Throughout her text, we see how she uses evidence to address different types of problems: getting to know her students at the beginning of the year, planning lessons, making decisions about what to do next in her interactions with students, and taking stock of their accomplishments. We also see how the ways in which she (and her students) use evidence shapes the nature of learning in that classroom.

Much of the evidence Lampert uses is a naturally occurring part of the (written and oral) discourse of the classroom. Some of it (quizzes and student notebooks) involves naturally occurring written records of students' work; much of it is simply present in the ongoing classroom dialogue. To learn about her students' learning from examples of their work, she examines their solutions and the strategies they used to reach them. Problem by problem, she considers the variations in students' performance, the types of errors they have made, keeps informal records of these in her journal, and designs her lessons accordingly. Lampert's journal—a routine part of her teaching practice—serves as an important space for jotting down evidence of interactions she wants to remember and for analyzing the evidence she has available. We also see how evaluation of work—by Lampert and her students—is a routine part of classroom interaction. Everyone participates in the assessment of the solutions proposed to problems. At the beginning of the year, we see the way in which she attends to students' performance on last year's standardized test. Instructively, she does not report (or seem to consider) their scores. Rather, she examines their work in solving the problems and considers, as she does with their notebooks, the approaches they took, the sorts of errors they made, and what these indicate about the meaning they make of mathematics. We see, as well, one example of how Lampert reports to parents about their student's progress—a prose description of what they have accomplished and what they have to work on—which is intended to represent the complexity of their mathematical performance in ways that grades cannot. It is important to note that Lampert routinely attends to evidence with a particular problem or question in mind about what to do next. Thus, evidence is always considered in light of the question or issue it is addressing: what problems to use in
tomorrow’s lesson, which mathematical concepts to illustrate with a given problem, what students’ likely approaches to a problem will be, which student to call on, how to respond to students’ misunderstandings, how to support students’ willingness to “study” and to question their own solutions, how to support students in being a resource for other students, how to communicate progress to students and their parents, and so on. Although we listed questions that seem to differentially emphasize the social and intellectual aspects of Lampert’s classroom, we note (as does she) that all teaching moves always entail both sorts of problems. And the evidence she records and uses is both about students’ evolving participation and identities and about how they solve problems.

It is also important to note that her attention to evidence in solving a teaching “problem” was routinely situated in the ongoing interaction of which it was a part, cumulative in the sense of drawing on other available evidence (fitting this piece of evidence into her evolving understanding of students’ learning), and anticipatory in the sense of considering how her next move was likely to affect students’ learning. Even the more formal quiz was evaluated in light of how it fit with other evidence of students’ progress, how the social situation and other affordances/ constraints of the quiz were different from the routine classroom interaction, and, equally important, how her response was likely to affect student’s sense of themselves as learners, their understanding of accomplishment, and their progress in her class. As her work with the quiz illustrates, her use of evidence always involved consideration of how her practice, including her assessment practice, shaped students learning.

What Expanding Conceptions of Assessment and Learning Suggest Validity Theory Needs to Accomplish

The theoretical perspectives and practices represented in this section illuminate the need for an expanded conception of validity to support the interpretive work in which teachers and other education professionals routinely engage. They show how evidence of learning is used to address different types of questions at different levels of scale; how many of these questions are fundamentally about what does and does not work to support learning and thus require evidence of the interaction between learners and their environment (including its assessment practices); how different conceptions of learning (e.g., evolving mental representations; evolving ways of participating in practice and the identities they entail) require different kinds of evidence of learning and opportunities to learn; how assessment practices range from discrete activities designated as assessment to ongoing ways of looking at the evidence available in the environment; how assessment practices always entail multiple interacting elements and function in complex systems that shape and are shaped by them. In short, the practice of assessment, and the IDAs it informs, are unavoidably situated in always partially unique contexts; a robust validity theory must be able to take these features into account.
VALIDITY THEORY AS SITUATED INQUIRY: RESOURCES FROM INTERPRETIVE SOCIAL SCIENCE

In this section, we describe and illustrate two sets of theoretical resources to support situated validity inquiry, both drawn from a constellation of perspectives that can be located within an interpretive approach to social science. Our goal is to develop an understanding of validity that begins with the questions that are being asked; that can develop, analyze, and integrate multiple types of evidence at different levels of scale; that is dynamic in the sense that questions, available evidence, and interpretations can evolve dialectically as inquirers learn from their inquiry; that allows attention to the antecedents and anticipated and actual consequences of their IDAs; and that situates the assessment in the broader context in which it is used.32

First, we draw on the tradition of hermeneutics to highlight general principles that might be considered in developing, investigating, documenting, and challenging the validity of situated IDAs in response to questions about educational practice. These tools are useful when validity claims, whether formal or informal, need to be made explicit. Second, we draw on the tradition of sociocultural studies to provide general guidelines for analyzing interactive, dynamic, and multidimensional educational environments. Analytical perspectives in sociocultural studies illuminate evidence relevant to the development and validity of situated IDAs and to the ways in which assessment shapes and is shaped by the education environments in which it is practiced. These sets of tools can be used both informally by actors within a system to reflect on their practice or formally by actors within and outside a system to better understand its resources and constraints for inquiry and learning.

It is important to note that these frameworks provide analytical perspectives (Lave & Wenger, 1991) that operate at a general level, much as the concepts provided by the educational measurement approach to validity do. We intend them to suggest types of questions that might be asked within any specific validity inquiry, but we argue that validity inquiries are always situated within a particular social context and guided by the problem, issue, or question one is trying to address and the available resources (evidence, conceptual tools) for addressing it. This illuminates the crucial role that concrete examples play in any representation of validity inquiry (much as they do in this chapter).33

Hermeneutics34

When the validity of an IDA in response to a question needs to be made explicit, whether formally or informally, the philosophical tradition of hermeneutics provides a useful set of analytic concepts. Like educational measurement, it provides a theory and practice of interpretation; it suggests validity questions, practices, and criteria that have important analogs in a range of approaches to interpretive social science or "qualitative" research methods35; and debates across fault lines within the tradition of hermeneutics (and between hermeneutics and educational measurement) pose instructive alternatives for validity inquirers to consider. Both educational measurement and
Hermeneutics provide means of combining information across multiple pieces of evidence, of dealing with disabling biases that readers may bring, and of bringing principles, criteria, or standards into contact with cases. Differences in how they address these problems can be framed, in part, in terms of the role that predetermined methods play in the development of an interpretation. Thus, hermeneutics provides a generative and theoretically coherent set of tools—coherent in the sense of creating a common language through which differences can be illuminated—to support multidisciplinary work.

We provide first a general overview of the tradition of hermeneutics, illuminating analytical concepts that can guide the development and evaluation of IDAs in response to questions. Then we turn to a more concrete example to illustrate the analytical perspectives.

An Introduction to Hermeneutics

Hermeneutics is about the theory and practice of interpretation, about the bringing of understanding into language. It was originally conceptualized as a practice for understanding written texts; however, the perceived relevance of hermeneutics has grown to include the interpretation of any meaningful social phenomenon, including complex (multitexted) social phenomena, such as historical traditions (see Bleicher, 1980, and Ormiston & Schrift, 1990, for introductions). Thus, hermeneutics suggests principles and issues relevant to the validity of IDAs based on “texts” or “text analogues” at quite different levels of scale, from a turn of talk or short answer essay to a recorded conversation or a portfolio of work to multimedia records of an implemented lesson or discussion of a test score report to compilations of multimedia records from a classroom or professional learning environment over time to accumulated scholarship in a field such as psychometrics or hermeneutics—all of which, of course, can be situated in larger social, cultural, and historical contexts. We will use the word “text” throughout this section to refer to this broad range of social phenomena of potential interest.

At the most general level, hermeneutics characterizes an integrative approach to combining sources of evidence in developing an interpretation. In this approach, readers seek to understand the “whole” body of evidence in light of its parts and the parts in light of the whole. Interpretations are repeatedly tested 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 referred to as the hermeneutic circle. As new sources of evidence are encountered, developed, and brought to bear, the hermeneutic circle expands, thus allowing a dynamic approach to interpretation.\textsuperscript{36}

Approaches to hermeneutics differ, however, in their representation of the role of the reader or interpreter. Here, we contrast “methodological hermeneutics” (as reflected, for instance, in the work of E. D. Hirsch) with the philosophical hermeneutics of H. G. Gadamer, on which we draw most heavily. In “methodological”
hermeneutics, the goal of the interpreter is to produce objective knowledge by grasping the “correct” meaning of text.37 Although the correct interpretation is typically associated with the intended meaning of the author of the text—a problematic criterion in the evaluative context of educational measurement—E. D. Hirsch (1976) argued that this methodology could be extended to evaluation. Hirsch conceived of evaluation as the accurate application of explicit criteria, a process that also assumed the author’s intended meaning was not distorted. The role of the interpreter, in both cases, was to bracket or control any prejudices they might bring to arrive at the correct interpretation or evaluation.

Viewed from the perspective of H. G. Gadamer’s “philosophical” hermeneutics (1987, 1994), the notion that a text has a correct meaning misunderstands understanding. Against this vision, Gadamer argues (following Heidegger) that: “There is no such thing as a fully transparent text” (Gadamer, 1981, p. 106), rather we always approach a text with fore-conceptions, presuppositions, or prejudices that are shaped by our history and that shape our understanding. Interpretations are unavoidably shaped by the linguistic and cultural resources the interpreter already possesses and by the nature of the questions the interpreter brings to the text (by why it draws the interpreter’s attention in the first place and by what the interpreter infers the text to be). In fact, without such preunderstandings, we would not be able to understand at all.

Thus, for Gadamer, the hermeneutic circle can be characterized as representing a dual dialectic: one between the parts of the text and the whole, and one between the text and the readers’ foreknowledge, preconceptions, or prejudices. However, if “unacknowledged presuppositions are always at work in our understanding” (1981, p. 111), if we need them to understand at all, then we must acknowledge that there are “legitimate prejudices” and ask “what distinguishes legitimate prejudices from the countless others which it is the undeniable task of critical reason to overcome?” (1994, p. 277). It is that question which undergirds Gadamer’s advice on developing interpretations.

For Gadamer, interpretation is most productively conceived as a conversation between two partners who are trying to come to an understanding about the subject matter in question. Gadamer (1987, 1994) offers some quite specific suggestions for the participants in a hermeneutic conversation that he characterizes as having a “hermeneutical attitude” (1987). He expects that all partners in the dialogue are willing to risk their own prejgments, to look for the coherence in what others are saying, and to believe that they have something to learn from the others. The “art” of conversation this requires is not the art of arguing against the other person but the art of “questioning” to bring out the strengths in the other’s argument (1994, p. 367). Thus, “reaching an understanding . . . means that a common language must first be worked out in the conversation” through which we can express our understanding of the other without distortion (p. 379, italics ours). “The important thing,” Gadamer (1994) tells us, “is to be aware of one’s own bias, so that the text can present itself in all its otherness and thus assert its own truth against one’s own meaning” (p. 269). Taylor (2002) suggests, following Gadamer, that better interpretations are those that
are less superficial, less ethnocentric and distorting, and more comprehensive. Taylor notes two important senses of the comprehensiveness of accounts: (1) "depending on how much detail and coverage they offer of the object studied" and (2) "on their taking in and making mutually comprehensible a wider band of perspectives" (p. 288, italics ours). Thus philosophical hermeneutic practices are productively conceived as a "conversation" (Gadamer, 1994) between an interpreter and a text and among interpreters about the subject matter in question.

This perspective has implications for the way criteria or principles should be brought into contact with texts. Given the importance of allowing what we are interpreting to challenge our preconceptions, Gadamer argues "the object itself must determine the method of its own access" (1987, p. 93). As Gadamer suggests (1987), the hermeneutic process used in making a legal judgment exemplifies the hermeneutic process as a whole. He argues that:

The judge does not simply 'apply' fixed, determinate laws to particular situations. Rather the judge must interpret and appropriate precedents and law to each new, particular situation. It is by virtue of such considered judgment that the meaning of the law and the meaning of the particular case are codetermined. (p. 148)

This does not deny the value of general principles or criteria. It does, however, change our understanding of their status and role. It suggests that the case plays a role in the criteria that are applied. It acknowledges that the criteria take on additional meaning when they are applied to a case, and that, rather than serving as "a guide for action," they become "a guide for reflection" (Gadamer, 1981, p. 82). Also it points to the importance of analyzed precedents (previously decided cases) in preparing interpreters to respond to new cases. When formal documentation is needed—as with consequential interpretations—the representation can take the form of an extended text. Interpreters can develop and represent interpretations that are appropriate to the available evidence and can situate the interpretations within descriptions of the evidence on which they are based and the process through which they were created. Remaining questions and insufficiencies in evidence can also be represented. This sort of representation allows others to evaluate the validity of the interpretation for themselves.

The discussion of hermeneutics, so far, suggests principles and raises issues about the development of interpretations that might inform validity theory. (1) The hermeneutic circle represents an integrative, dynamic process where questions, evidence, and interpretations are iteratively elaborated and revised. This allows the development of unique interpretations in response to unique configurations of evidence. (2) Different theoretical approaches to hermeneutics provide alternative and partially competing perspectives on the role of a priori criteria and the stance that readers might take. These include: (a) seeking to develop the correct interpretation or application of a priori criteria or (b) allowing the text to codetermine the criteria through which it is evaluated. (3) Seeking challenges to preliminary interpretations, from other interpreters and the text, is a crucial means through which sound interpretations are developed. (4) Learning from precedents—interpreted texts—as well as from abstract principles
plays a crucial role in developing sound interpretations. (5) Agreement between interpreters and text, and among interpreters, who have actively sought challenges to their preconceptions in the others' perspectives, is a criterion of a good interpretation.

Scholars of Critical Theory criticize Gadamer's philosophical hermeneutics for failing to acknowledge and address the limits of hermeneutic understanding and point us to what is often called "critical" or "depth" hermeneutics. Habermas argued that hermeneutic consciousness is "inadequate in the face of systematically distorted communication" (1990b, p. 253), which occurs when participants in a dialogue are influenced by social, political or economic forces of which they are unaware. This can result in a "false consensus." His proposed corrective can be seen to involve two major components.

First, he offered us a regulative ideal (that he acknowledged is rarely achieved in practice): Interpretations are only valid if they are reached because of the power of good reasons, if they are unaffected by coercion or the effects of unequal power, and if everyone has an equal chance to contribute to the dialogue (see, e.g., 1990c, pp. 88–89; 1993, p. 164). Habermas noted that when interpretations or decisions are intended to regulate people's actions or access to resources, interpreters have an added obligation: in addition to the observational evidence, they should consider the "consequences and the side effects . . . [that it] . . . can be anticipated to have for the satisfaction of everyone's interests" (1990c, p. 65).

Second, Habermas (1990b) argued that what is needed is a critically enlightened or depth hermeneutic "that differentiates between insight and delusion" (p. 267) in the process of discourse. Here, the goal is to explain the text and the interpretations of it in terms of the conditions that produced them. Studies that do this can range from large-scale studies to microanalytic studies of interaction that seek to connect meaning/interpretation to the social structures that shaped it.

Thus, critical or depth hermeneutics highlights the role that readers from outside the interpretive community can play—bringing an alternative perspective that illuminates the values and theories taken for granted by those within the interpretive community, so that they may be self-consciously considered. However, Habermas (1990b) noted, "there is no validation of depth-hermeneutical interpretation outside of the self-reflection of all participants that is successfully achieved in a dialogue" (p. 270).

Again, the discussion suggests principles and raises issues about the development and evaluation of interpretations that contribute to validity theory. (1) Again we see the privileging of agreement or consensus as a criterion of a sound interpretation, although Gadamer criticized Habermas's ideal of uncoerced agreement as "empty" (1990, p. 293). (2) The difference between the third-person explanatory discourse of depth hermeneutics and the first- and second-person dialogue of philosophical hermeneutics offers interpreters productive alternatives that may be dialectically engaged. (3) The principle of depth interpretation—outside critical analysis of the interpretation and the process through which it was achieved—is useful in illuminating the social influences that may be operating beneath the
interpreters' consciousness. (4) Consideration of consequences is important, especially for interpretations that control people's actions or limit their access to society's goods.

Scholars of more pluralist approaches to hermeneutics criticize both Habermas and Gadamer for seeming to privilege agreement or consensus among interpreters and the texts they are interpreting as an important criterion for a good interpretation. These theorists suggest that the goal of a successful hermeneutic conversation is to understand and learn from our differences. As Hoy (1994) described it: "The hermeneutic model calls for enlarging one's interpretations and enriching them by holding them open to other interpretations" (p. 264). Although agreement may be a welcome side effect: "interpreters can believe that their understanding is reasonable and right without also believing that everyone else will or even should agree with them" (1994, p. 182). The point, suggested Warnke (1994), is for participants "to be sure their own interpretations are as compelling and inclusive as they can be" (p. 133). From this perspective, a hermeneutic interpretive conversation becomes "a way to clarify our disagreements" (Bernstein, 1992, p. 338; see Moss & Schutz, 2001, for a fuller discussion), and it is an opportunity for the interpreters' learning.

That said, there are times when consequential decisions need to be made and actions taken. Consequential decisions and actions bring an arbitrary closure to the process of interpretation, which is otherwise open to new evidence and perspectives. Also, consequential decisions typically do not permit multiple resolutions; they require that a single best course of action be determined and justified. Sometimes there are a priori limitations on the available courses of action (e.g., a selection or placement decision; a decision about certification or licensure); sometimes unique and complex resolutions can be constructed (e.g., in the design of learning opportunities for students or professionals; in funding decisions).

Hermeneutics raises numerous productive questions relevant to validity that might be asked about the process and outcome of decision making: (1) What criteria and evidence can be used in making a decision, and when, how, and by whom they should be determined (e.g., do they remain open during the decision-making process, such that new sources of evidence and criteria can be considered, or are they specified before the process begins)? (2) Who participates in the decision-making process, what roles do they play, what affordances and constraints does each have for influencing the decision, and what stances does each take vis-à-vis the evidence and one another? (3) What role should anticipation and evaluation of the consequences of the decision play? and (4) Should (and how should) the process be evaluated to illuminate potential problems about (a) how ambiguous or insufficient evidence is handled, (b) how disagreements are resolved, (c) how differences in power and status shapes participation, (d) how experience and knowledge—in general and about the local context—shape participation, and (e) how preconceptions, biases or ideologies are illuminated and addressed? We note that these questions might equally well be asked about standardized decision-making processes as developed in educational measurement and that the questions illuminate important differences in the
procedures. They point to the ways in which validity, or soundness of IDAs, cannot be separated from issues of ethics and power to shape them and their consequences.

**An Illustration of an Interpretive Practice Informed by Hermeneutics**

To illustrate the meaning and potential of the sometimes competing principles outlined previously, we offer the following concrete example based on our own work with the use of teaching portfolios to inform licensure decisions (Haniford & Girard, 2006; Moss & Schutz, 2001; Moss et al., 2004; Moss, Cogshall, & Schutz, 2006; Schutz & Moss, 2004). Although the assessment systems with which we have been working have been guided largely by the principles of educational measurement, our validity research has drawn on hermeneutic practices to illuminate assumptions and consequences and to highlight potential problems of the more conventional practices.

We focus here on a special study where small groups comprising experienced math teachers, math teacher educators, and mathematicians were asked to review portfolios submitted by beginning secondary mathematics teachers with the question of whether these teachers should be granted a renewable license. The portfolios had originally been submitted by beginning teachers to their state’s department of education—passing this standards-based portfolio assessment is required to convert an initial license to a renewable license. Teachers are asked to include the following kinds of evidence in their portfolios: a description of the classroom context, descriptions of a series of lessons with instructional artifacts (e.g., handouts, assignments); videotapes, student work, and reflections on two featured lessons; a cumulative evaluation of student learning with accompanying reflection; a focus on two students throughout the featured lessons and cumulative evaluation of learning; and analysis of teaching and professional growth. We asked our readers to try to reach a similar decision but without the constraints and supports of the state’s assessment system (which specified criteria and features of the portfolio to which readers were allowed to attend). Thus, our readers had a circumscribed evidence base but were asked to evaluate the portfolios in terms of their own criteria. First, they wrote independent narrative evaluations of the portfolio, along with a justification for the decision they had made; then they came together to debate the decision for those portfolios on which they had disagreed. Readers can find our extended analysis in Moss, Cogshall, and Schutz (2006).

In both the discussion and the independent written interpretations provided by the readers, we see a dialectical process of reference to pieces of evidence that raised questions that caused them to seek additional pieces of evidence and so on. This led to interpretations qualified in the available evidence (and the acknowledgment where evidence seemed to be insufficient or ambiguous) and the request for additional information where questions could not be addressed. Consider, for instance, their discussion of the quality of teaching in two portfolios and the cues they drew on to evaluate it. In a portfolio from an advanced mathematics teacher, they considered that the teacher had designed lessons that went beyond the mathematics in the text book, that
she was willing to let her students struggle with a difficult problem, that the mathematical discourse showed good "sensibility" for mathematical issues, that she was sometimes sloppy in her mathematical language, that one of the lessons ultimately failed to find solutions to the difficult problem that had been set, and that she was not able to articulate connections between her teaching and what students were learning. With a portfolio from a teacher of introductory algebra, they observed a teacher who was communicating clearly about the mathematics but working directly from the textbook, in what appeared as a procedural approach. They also noted that the teacher's description of his practice did not match the videotape and other artifacts available in the portfolio and that his reflections illuminated an emotional response to the students and his own teaching that seemed problematic (in that it did not seem to allow him to explore productive solutions to the problems he perceived).

Central, as well, to their determination of an ethical decision was their estimate of the teacher's potential. One way they attempted to estimate a candidate's potential was by seeking to understand more about the context in which the candidate was working and the type of support he or she had received: for instance, with both portfolios, they wondered about what and how they were required to teach, what opportunities for learning their preparation program had provided, whether their choices were shaped by what they perceived were the needs of the portfolio readers, and so on. In many cases, these were questions that could not be resolved with the available evidence. With both portfolios, they argued about different possible explanations for the performance, which might imply more or less potential. In addition to context, they looked for other cues to help determine the teachers' potential: what they had contributed to the curriculum they were teaching, whether their subject matter knowledge provided a sufficient base on which to build, whether they could analyze their teaching in a way that would lead to productive changes, and so on. Again, the evidence in the portfolio was typically insufficient (and the portfolio was not designed) to answer these questions.

The readers uniformly resisted the limited interpretations and consequences that were available to them within the state's system. As one reader noted: "[E]ven though we may give two different people a two level, the reasons for the 'two' [pass] are highly differentiated. . . . These people each have quite specific and particular kinds of problems and they differ from one case to another." For each of the portfolios that were debated, the readers concluded that the teachers needed additional professional development of various sorts (consistent with what would happen for second-year teachers who failed) but were unwilling to arrive at a failing decision (in case these were third-year teachers who would not be allowed to continue) because there was either promising or insufficient evidence to determine the teachers' potential. They worried about various potential collateral consequences of providing teachers with a score—that low scores would have a negative psychological effect, that the system may not have a sufficient number of teachers to staff its schools, that a passing score would send the message that no additional work needed to be done, or that a high score would encourage wealthy districts to skim off the best teachers. They could
not find a comfortable decision within the options that the state provided. Instead of distinguishing between more and less effective teachers and determining who would be licensed and who would not—which is while some limited professional development—the expert readers argued that the main focus of a portfolio assessment process should be on professional development. Of course, the actual consequences of the decisions in terms of the issues the readers have raised are empirical questions about how the assessment practice functions within the larger system, which sociocultural studies will help address.

Clearly, the given example does not illuminate all of the principles that might be considered in developing an interpretation. However, it does illustrate a dialectical, dialogical process of developing questions, evidence, and interpretations in attempting to reach an ethical decision. It also points to the deeply situated nature of interpretations and their consequences. It suggests the importance of understanding multiple elements of the activity systems (1) in which the evidence was produced (e.g., the readers’ questions about the context of the teachers preparation and work), (2) in which it is being evaluated (what enabled and constrained the readers’ interpretations), and (3) in which its consequences will be experienced as part of validity inquiry. Again, sociocultural studies provide theoretical resources that address these issues directly.

A SOCIOCULTURAL FRAMEWORK FOR ANALYZING LEARNING ENVIRONMENTS

The practice of assessment always happens within a particular activity system, community of practice, or learning environment. As illustrated by our previous discussion, questions about learning and about how assessment documents and supports (or impedes) learning entail an understanding of the entire activity system. In an earlier section, we focused briefly on sociocultural understandings of learning and their implication for assessment. Here, we focus on the rich potential of sociocultural studies for guiding analysis of complex activity systems or learning environments. We refer here to any activity system in which evidence is produced and used and to all the actors (including education professionals) who work and learn within them. We focus first on providing conceptual tools that are useful in analyzing a single “focal” activity system or learning environment. Then we turn to the situation where information crosses boundaries to and from an external activity system that is responsible for supporting and monitoring the focal activity systems. Our goals are to suggest (1) the range of sources of evidence that might be considered in supporting/challenging IDAs about learning, (2) analytical perspectives through which we might examine how assessment practices function in a learning environment, and (3) an approach to validity inquiry that examines the learning of the professionals who are using the assessments.

We use the term “sociocultural” to encompass a range of theoretical perspectives, which, as Lave (1993) described, share an interest in “relations among the person,
activity, and situation, as they are given in social practice” (p. 7). We draw in particular on three sets of resources: cultural-historical activity theory (activity theory or CHAT, for short) (Engeström, 1987; Engeström, Miettinen, & Punämaeki, 1999), the sociocultural theory of mediated action (Wertsch, 1998; Wertsch, Del Rio, & Alvarez, 1995), and a situated theory of learning (Lave, 1988; Lave & Wenger, 1991; Wenger, 1998).41

Elements of a Learning Environment, Community of Practice, or Activity System

As Wertsch (1998) described, “The goal of a sociocultural approach is to explicate the relationships between human action, on the one hand, and the cultural, institution, and historical situations in which this action occurs, on the other” (p. 24). Human action includes human cognition and much more: “Action may be external as well as internal, and it may be carried out by groups, both small and large, or by individuals” (1998, p. 23). Wertsch (1998) argued that we need to treat “mediated action” (p. 17) or “individual-operating-with mediational-means” (p. 26) as a primary unit of analysis. Mediational means are both physical (e.g., texts, calculators, and measuring instruments) and symbolic (e.g., concepts, language systems, representational schemes, and standards). They include the cultural tools that a community inherits and adapts, as well as those they produce and evolve. Cultural tools “provide the link or bridge between the concrete actions carried out by individuals and groups, on the one hand, and cultural, institutional, and historical settings, on the other” (Wertsch et al., 1995, p. 21).

Others foreground the entire community of practice (Lave & Wenger, 1991) or activity system (Engeström, 1987, 1993, 2001), including conceptual and physical tools and other people as the unit of analysis. For Lave and Wenger, “a community of practice is a set of relations among persons, activity, and world, over time and in relation with other tangential and overlapping communities of practice” (p. 98). With Wertsch, they noted that tools or “artifacts—physical, linguistic, and symbolic” (p. 57) are particularly important because they carry a community’s heritage. Artifacts and social structures “leave a historical trace . . . which constitute and reconstitute the practice over time” (pp. 57–58).

Similarly, for Engeström, an activity system builds on the concept of mediated activity (or individual acting with mediational means) to “explicate the societal and collaborative nature” of actions (1999, p. 30). Engeström depicted the components of an activity system as follows.

In the model, the subject refers to the individual or subgroup whose agency is chosen as the point of view in the analysis. The object refers to the “raw material” or “problem space” at which the activity is directed and which is molded or transformed into outcomes with the help of physical and symbolic, external and internal tools (mediating instruments and signs). The community comprises multiple individuals and/or subgroups who share the same general object. The division of labor refers to both the horizontal division of tasks between the members of the community and the vertical division of power and status. Finally the rules refer to the explicit and implicit regulations, norms and conventions that constrain actions and interactions within the activity system. (Engeström, 1993, p. 67; see also, Engeström, 1987, for a fuller explication)
As Engeström noted, activity systems are neither static nor closed. They involve ongoing relationships among people and their world where the elements are continually reproduced and/or transformed. Thus, it is important to understand how an activity system has evolved over time, in terms of both its local history and the "global" history of the concepts, procedures, and tools it inherits (Engeström, 1999, p. 137). Indeed, from a sociocultural perspective, learning is perceived and understood in terms of evolving relationships among learners and the other elements of their learning environments (Lave & Wenger, 1991, p. 51). Transformations (or system-level learning) often happen when systems encounter contradictions, when new elements are introduced, or when alternative perspectives and practices are encountered. (The value of alternative perspectives and the challenges they raise is a theme encountered in educational measurement, as well as hermeneutic approaches to interpretation.)

A crucial element for our analysis is the understanding that learning is a central component of any activity system, whether or not explicitly intended to foster learning. Questions about what is being learned are equally relevant to activity systems where educational professionals work with one another as they are to classrooms. As Lave (1993) stated: "learning is an integral aspect of activity in and with the world
at all times” (p. 8). It entails a dialectical relationship between experience and understanding. From this perspective, learning does not just happen as a response to teaching (where, they note what is learned may be quite different from what is intended to be taught); rather learning happens everywhere, all the time, as we participate in social (inter)action. A learning curriculum unfolds in opportunities for engagement in practice (Lave & Wenger, 1991, p. 57). Furthermore, learning always involves the “construction of identities”:

In this view, learning only partly—and often incidentally—implies becoming able to be involved in new activities, to perform new tasks and functions, to master new understandings. Activities, tasks, functions, and understandings do not exist in isolation; they are part of broader systems of relations in which they have meaning. . . . Learning thus implies becoming a different person with respect to the possibilities enabled by these systems of relations. (Lave & Wenger, 1991, p. 53)

With respect to evidence use in and outside the classroom, this suggests we need to illuminate and analyze the “learning curriculum”—the opportunities for engagement in practice and the types of knowledgeable identities that different approaches to assessment afford teachers, administrators, policy makers, and other professionals, as well as students. Furthermore, we need to understand how people can learn when there is no explicit curriculum to point the way. Engeström’s (1987, 2001) conception of “expansive learning,” which characterizes the kind of learning that productive organizations engage in when “we must learn from new forms of activity that are not yet there” (2001, p. 138).

In analyzing an activity system, Engeström (1999) suggested, “it is fruitful to move from the analysis of individual actions to the analysis of their broader activity context and back again” (p. 32). Lave and Wenger (1991) suggested that the analysis of a community of practice would involve the following types of questions:

• about the sociocultural organization of space into places of activity and the circulation of knowledgeable skill;
• about the structure of access of learners to ongoing activity and the transparency of technology, social relations, and forms of activity;
• about the segmentation, distribution, and coordination of participation and the legitimacy of partial, increasing, changing participation within a community; [and]
• about its characteristic conflicts, interests, common meanings, and intersecting interpretations and the motivation of all persons vis-à-vis their changing participation and identities (pp. 55–56, paragraph marks and bullets inserted).

Lemke (2000), a frequent dialogue partner with our focal theorists, provided a useful elaboration for our conceptual toolkit, which focuses on levels of timescale in analyzing processes in dynamic “ecosocial” systems of interdependent processes. Lemke suggested that when analyzing ecosocial systems (of which the educational system is an example) there are two fundamental questions: “What processes, what kinds of change or doing, are characteristic of each relevant timescale of organization of the system/network? and How are processes integrated across different timescales?” (p. 275). He
suggested further that it is useful to analyze scale hierarchies in groups of three levels at once. For any given activity (happening or doing), this suggests the importance of looking at both what happens before and after and one level up and down in a hierarchical timescale.

For adjacent timescales it is also quite clear that the processes at the next lower timescale make possible the repeatable patterning of the next longer scale. . . . What is equally important, however, is that there is always also a higher level process already in place, already running on its own longer timescale, and this sets the context that constrains what is likely and what is socially appropriate at the next scale below. (p. 276)

For instance, the activity of the quiz in Lampert’s classroom, which unfolds over a relatively short period of time, comprises a series of briefer interactions among the students, the instructions Lampert gives them to respond to, the responses they produce, the conceptual resources they use in responding, and Lampert’s oral and written feedback on their responses. It is also situated in an implemented learning curriculum that gives students conceptual resources, roles, and responsibilities, and so on. It takes on meaning against this background.

These elements of an activity system are general analytic concepts that suggest questions to be asked about any group of people working together, with conceptual and physical tools, around an object. They are not intended, in and of themselves, as a description of a social reality. The full characterization of any particular activity system is typically constructed through empirical work involving observations, interviews, artifact analyses, and so on. It represents a generalized pattern that can be constructed, over time, from the study of particular (mediated) actions and interactions.

Thus, the conceptual tools we described call attention to people who take on different identities and positions (with respect to one another and the conceptual and physical tools); the conceptual and physical tools they inherit, adapt, or produce and use; the norms and routines in which they engage; the objects and intended outcomes that focus and motivate their actions; and the other communities of practice or activity systems with which they interact. They suggest that analysis of any happening must consider what happened before and after, the briefer happenings (e.g., actions and interactions) of which it is comprised, and the longer happenings of which it is a part.

**Examples of Evidence Use in Different Activity Systems**

To illustrate how tests and other sorts of evidence of learning and teaching have been incorporated into activity systems at the school level and how these practices might be analyzed with the types of resources described, we draw on examples from the work of James Spillane and colleagues with the Distributed Leadership Study. Although Spillane’s focus was on “leadership practice,” we can nevertheless read his work for how assessment and other evidence of teaching and learning function. The examples we cite involve the use of evidence from the classroom, of one sort or another, that crosses the classroom boundary to be used by teachers and administrators in
school level activity systems (Diamond & Spillane, 2004; Spillane, 2006; Spillane, Halverson, & Diamond, 2001; Spillane, Diamond, & Halverson, 2004).

Distributed Leadership Study researchers examine how the relationship between leadership activities and teacher’s classroom work is shaped by various institutional/social structures in which they are embedded and by conceptual and material tools that mediate those relationships. Material resources or tools include artifacts such as students’ tests and test scores, curriculum guides, text books, other printed materials, Internet and other technology resources, observation protocols, state and district standards, forms, and meeting agendas. These tools “mediate” (shape, enable, and constrain) leaders’ actions; the sense leaders make of them in turn mediates the effect of the tool. Similarly, leadership practice is shaped by conceptual or cultural tools that leaders use to make sense of the ideas they encounter. These conceptual resources include “language, theories of action, and interpretive schema” that enable “intelligent social activity” (Spillane et al., 2001, p. 23): “even when a particular cognitive task is undertaken by an individual apparently in solo, the individual relies on various sociocultural artifacts, such as computational methods and language that are social in origin” (Spillane et al., 2001, p. 23, citing Wertsch).

Leadership activities occur within institutional and social structures that also shape them. These include structures (formal and informal relationships and routines) that have been developed within the school, such as subject matter or grade-level departments, faculty meetings or classroom observations, and time and space set aside for teachers to plan together or the lack of such opportunities. They also include the many external structures in which schools and their leaders are embedded, such as district and state education agencies, professional organizations, legislative requirements, school-community activities, and so on. These are all contexts and routines that bring people together in various configurations for various reasons. They shape the way leadership is practiced and how conceptual and material resources, such as tests and test scores, are used.

Distributed Leadership Study publications are rich with examples of how formal and informal evidence of teaching and learning is used by teachers and administrators in different social arrangements. Below we describe a series of vignettes, culled from this research agenda with 13 schools for 4 years, that draw on interviews, observations, and surveys.

Vignette 1: Teaching Observations

For instance, Spillane and colleagues (2004) describe how the instructional evaluation is carried out, interdependently, by the principal and assistant principal at one school:

The assistant principal, who maintains a friendly and supportive relationship with teachers, visits classrooms frequently and engages in formative evaluation by providing regular feedback to teachers on instructional issues. The principal, on the other hand, functions more as an authority figure and engages in summative evaluation. She visits the classrooms one to two times per year and makes final determina-
tions on the quality of teachers’ instructional practices. The assistant principal shares his learning with the principal, and the two use their collective observations to develop an understanding of teachers’ instructional practices. (p. 17)

As the researchers noted, even though the two leaders appear to work separately, the evaluation practice at the school can only be understood by considering both practices. The same activities undertaken by either alone might amount to a different practice with different effects. To illustrate the importance of evaluation artifacts in the shaping of this activity, they invited readers to consider how the activity would differ based on two different protocols: focused on (1) a checklist of generic teaching processes (such as use of wait time and praise) or (2) questions about subject-specific practices, such as how mathematical tasks are represented or how students are required to justify their mathematical ideas. As they noted, these different forms draw observers’ attention to different aspects of the teaching situation, and, thus, the leadership practice is likely to promote different kinds of teaching practice.

**Vignette 2: Students’ Writing as Evidence**

For an example of how another source of evidence can shape leadership practice, the principal at one school engages in monthly reviews of writing folders from each classroom. The writing folders contain a sample of writing from each student in a teacher’s classroom. The principal writes comments to the teacher on each folder and to each student. She keeps copies of her comments to the teachers in the file, monitors progress relevant to the comments in subsequent writing folders, and considers this in each teacher’s summative evaluation at the end of the year. Interviews with teachers suggest how this leadership practice has shaped their teaching practice. One teacher reported, for instance, that she had switched her whole day around so her students got an hour for writing (Spillane, 2006). To illuminate the importance of different tools, Spillane asked his readers to consider what would have happened if the principal had decided to review teachers’ lesson plans instead of their students’ work.

In a contrasting example using students’ written work (Spillane et al., 2004), a teacher leader and assistant principle held biweekly professional development meetings “to foster reflective dialogue” that would help teachers in facilitating writers’ workshops in their reading classrooms. Teachers shared their accounts of how they enacted “Writer’s Workshop,” as well as examples of the stories their students had composed. Here the focus was on informal use of evidence to support collaborative learning among teachers to help improve their reading instruction. Again we see how differences in even one element of an activity system can shape the practice.

**Vignette 3: External Test Scores as Evidence**

Diamond and Spillane (2004) contrasted the use of test-based evidence in four schools, two that had been placed on probation and were at risk of being restructured and two that were relatively high performing on the district-mandated tests. They
noted that leadership practices in all four schools showed evidence of attention to test scores, including prioritizing the subject areas and content covered and providing some form of explicit test-preparation activities for students. However, they observed important differences in how leaders from the different schools interpreted test-based information and used it to inform instruction. The high-performing schools used test results to both identify trends and focus attention on areas of specific need (p. 1164). At the probation schools, they saw less focus on instructional implications: “School leaders discussed the need to improve reading and mathematics and did speak in specific terms about subdimensions of these subject areas but did not speak about specific instructional approaches and strategies as they did in the high performing schools” (p. 1165). Diamond and Spillane described probation school leaders working to manage the impression of outsiders, tending to focus more on the appearance of instructional innovation than substantive change. They also noted a tendency to “target” students and grade levels “to increase the number of students at or above cutoff points at benchmark grades” (p. 1155). They noted that in the higher performing schools: “the resources to enable data interpretation are higher” (p. 1165), staff work together to analyze the information provided to define specific instructional needs that provide a basis for instructional decisions, and they are, therefore, more likely to benefit from the information.

Thus, in the vignettes of leadership practices at these schools, we see examples of how formal and informal evidence of teaching and learning is used by teachers and administrators in different social arrangements. We see different types of evidence, different resources for interpreting them, different divisions of labor and positioning of adult learners, and different understandings of the appropriate outcome. Again, the vignettes point to the importance of understanding how assessment functions as part of a complex activity system.

**Extending Our Analytic Framework to Include Assessments That Cross Boundaries**

In this section, we focus on analytic questions that might be asked about evidence that crosses boundaries between activity systems either formally (systematically documented) or informally and with and without people to recontextualize it. Although this includes the large-scale student testing programs that states and districts use, it also includes the many (formal and informal) boundary crossings illustrated in the vignettes from the Distributed Leadership Project described previously. Thus, how evidence crosses boundaries varies on multiple dimensions, including the (1) types of evidence (narratives of practice, samples of student work, videotapes of classroom interaction, interviews and observations, and surveys, as well as standardized tests used by districts and states), (2) the extent to which the evidence is explicitly and systematically documented, (3) whether and how actors from the local context accompany the evidence into the new context and can thus recontextualize it, (4) the comprehensiveness of the representation of the local context (what is made visible and what
remains in the background), (5) what norms and routines accompany the providing and interpretation of evidence, and (6) what roles the actors in the local and external contexts play in shaping the way in which practice is represented.

When evidence crosses boundaries, it brings far more than information: it entails sets of cultural tools, including artifacts, concepts, and often norms and routines, that mediate understanding and (inter)action in sending and receiving contexts. In the case of large-scale standardized assessments, for instance, the set of cultural tools includes artifacts, such as stated goals of the assessment, test forms, standards or domain descriptions, guidelines for evaluating performances, score reports, technical manuals, regulations for users; it includes concepts that represent what is important to learn and what counts as evidence of that learning; it includes expected practices (rules and norms), such as standardized administration formats, independent work; and it entails an implied division of labor (different roles for test developers, teachers, students, parents, administrators, policy makers, and others) in the construction and use of the information. In Greeno’s terms (2002), it positions the different actors with differential authority, accountability, and agency for making decisions about how their practice is represented and how those representations should be interpreted and used. In short, it provides a partial vision of an activity system through which the assessment is enacted.

Therefore, it is important to remember that externally mandated tests are always interpreted and used in particular local contexts, which shape and are shaped by them. Depending on how they are implemented by the central authority—that is, what practices are required/expected to accompany them—and how they are taken up in the local context, they may involve more or less incentive to conform to their particular vision of learning. As Wertsch and colleagues (1995) note: “While cultural tools shape action, they do not determine it” (p. 22). Furthermore, when individuals use the tools in particular settings, both the tools and the setting [or activity system] are transformed (p. 26). Thus, “mediation is best thought of as a process involving the potential of cultural tools to shape action, on the one hand, and the unique use of these tools, on the other” (p. 22) (italics ours). As we have seen, the same tools can be taken up in different ways in different environments, to different effects. To the extent that externally mandated assessments are used across many local contexts, they can have a powerful and widespread effect, for better or worse (Bowker & Star, 1999; Jordan & Putz, 2004). Understanding this effect is an important element of validity inquiry, in no small part, because it shapes the learning that the assessments are intended to document.

The concept of a “boundary object” or “boundary infrastructure,” developed by Star and colleagues (Bowker & Star, 1999; Star & Griesemer, 1989) provides additional theoretical resources for the analysis of assessments that cross boundaries. A boundary object is an object that inhabits multiple heterogeneous social worlds (or activity systems) and that enables communication and cooperation across these worlds. “Boundary infrastructures” involve “objects that cross larger levels of scale than boundary objects” (Bowker & Star, 1999, p. 287) as is typical with centrally
mandated assessments. As they note, participants from different social worlds each "answers to a different set of audiences and pursues a different set of tasks" (p. 388) and "because . . . objects and methods mean different things in different worlds, actors are faced with the task of reconciling these meanings if they wish to cooperate" (p. 388). Furthermore, "unless they use coercion, each translator must maintain the integrity of the interests of the other audiences in order to retain them as allies" (Star & Griesemer, 1989, p. 389).

Thus, a boundary object is a particular kind of cultural tool that not only crosses boundaries of activity systems, such as mandated assessments, but also is plastic enough to adapt to local needs while maintaining a common identity across sites (Star & Griesemer, 1989, p. 393). It enables translation and, therefore, cooperation, but without coercion. A mandated assessment would function as a boundary object when actors in the local context are able to cooperate in providing necessary information to outsiders while maintaining a productive level of authority and agency over their own practice. Star and Griesemer note that the function of boundary objects cannot be understood from a single perspective. Rather, it requires an ecological analysis that examines both local (situated) and shared meanings, that traces the function of the boundary object across worlds and does not, therefore, privilege a single point of view:

The major requirements, for such an ecological understanding of the path of rerepresentation, are thus: (1) How objects can inhabit multiple contexts at once, and have both local and shared meaning. (2) How people, who live in one community and draw their meanings from people and objects situated there, may communicate with those inhabiting another. (3) How relationships form between (1) and (2) above—how can we model the information ecology of people and things across multiple communities? (4) What range of solutions to these three questions is possible and what moral and political consequences attend each of them? (Bowker & Star, 1999, p. 293)

Thus, with externally mandated assessments, ideally and eventually, one would want to analyze (examples of) all the activity systems in which the assessment functions (how it shapes and is shaped by the local practice): this would include the activity systems through which the assessment was conceptualized, developed, mandated, and implemented; the school and classroom activity systems in which it is responded to, interpreted, and used; the activity systems involving administrators and policy makers at the district, state, and national levels; the activity systems of students' families and peer groups; the activity systems of professional organizations and teacher education institutions that attend to such information; and the "virtual" activity systems of members of the public who attend to evidence about how their educational systems are functioning.

Using Evidence That Crosses Boundaries: An Extended Example

The Data Wise inquiry model (Boudett, City, & Murnane, 2005), developed by researchers and graduate students at Harvard in conjunction with teachers and administrators in the Boston Public School system, provides a rich example of how
externally mandated tests can be used as part of a local inquiry process. It illustrates, in our judgment, the potential for external tests to serve as boundary objects. The model consists of three basic phases (Prepare, Inquire, and Act), which are further subdivided into eight steps. In what follows, we summarize this process, highlighting the elements that represent concrete instantiations relevant to our argument.

The “Prepare” phase consists of building a foundation for data-driven inquiry in a school community by creating a community of adult learners. They note that school improvement efforts are likely to be more effective if responsibility for data interpretation is shared among school community members. Data Wise also suggests that the members must have “assessment literacy”—an understanding of how to read and interpret standardized test score reports. Although this can be initially supported by outsiders, such knowledge and capacity needs to become part of the larger community’s repertoire.

As part of the initial phase, the Data Wise team suggested that schools create a “data inventory” of the types of data that are present in their system as a first step in the use of formal inquiry into school data. Examples of common data sources available in schools include standardized test results (including both state- and district-level tests), developmental reading assessments, observation surveys, running records, writing samples, unit assessments, other student background information (e.g., ethnicity and language proficiency), and attendance records. The inventory distinguishes between internal and external assessments (e.g., running records vs. state skill mastery tests), and then for each assessment suggests the following categories of information be included: content area, dates of collection, which students are assessed with it, who has access to the data, how it is currently used, and the potential for more effective uses (p. 15).

The next phase is “Inquire,” which consists of three steps. First, schools must get a firm grip on the data that they have at their disposal already, which has been outlined in the data inventory discussed. Other types of data that might be developed include, for example, artifacts from classroom practice (such as class work and homework), student interviews, and teacher peer observations. Next is “Digging Into the Data,” and it is at this interpretive phase that the advice provided becomes more clearly connected to our hermeneutic concepts. For example, Data Wise makes explicit the potential and value of data, like student work, to challenge teachers’ assumptions, a key element of the hermeneutic process:

Examining student work helps to surface and challenge many assumptions—assumptions about what students can and cannot do, about which students can do what, and about why students are or are not able to do something. Challenging these assumptions is important for three reasons. First, you want the clearest understanding possible about the student learning problem, and assumptions often obscure this understanding by taking the place of evidence. Second, teachers fundamentally have to believe that students are capable of something different from the results of the current data. Otherwise, why bother putting any effort into helping students learn? And third, the solutions for the problem will require changes in what the faculty members do on a day-to-day basis. Making significant changes in what you do often requires changing what you believe. Opportunities for teachers to share their interpretations of student data provide occasions to address these fundamental beliefs about learning and teaching. (p. 88)
"Examining Instruction" is one of the explicit steps in the inquire phase, and attending to the particulars of the classroom is an important source of evidence for such inquiry. Accordingly, Data Wise places the examination of current practice at the heart of its investigation process.

Next, the Data Wise "Act" phase involves planning, action, and assessment. Of note is that in parallel with the development of an action plan is an emphasis on assessing the action, which creates a new corpus of data with which to start the inquiry cycle again. Furthermore, the criteria for success of the action plan may include both standardized assessments, as well as "home grown" measures determined by the inquiry team.

Like Data Wise, there are multiple projects that involve schools and districts in inquiry-based models of reform or organizational learning that combine many of the elements described above. Additional examples are listed in an endnote. These models vary, instructively, in the types of evidence considered, the processes that surround its development and interpretation, the extent to which the evidence is comparable and aggregatable across contexts, the roles actors in local contexts play in representing their own practice to those to whom they are accountable, and so on.

We note that the focus of our examples has been on evidence use at the classroom, teacher community, and school level, although we argue that these analytical perspectives suggest categories of questions that might be asked of communities of practice or activity systems at different levels of the activity system.

CONCLUSION

Our goal in this chapter has been to sketch and illustrate a constellation of theoretical perspectives that might be used to conceptualize the validity of evidence-based interpretations, decisions, and actions routinely made by teachers and other educational professionals. We see this as a preliminary step in what we believe should be an ongoing multidisciplinary agenda of research and practice. The theoretical perspectives we have represented suggest categories of questions that might be asked about the use of evidence in any learning environment, including those in which education professionals are learning to support students' learning and one another's learning (Moss & Greeno, in press). We have pointed to examples of existing studies and representations of practice that might be situated within such an agenda. However, as these examples demonstrate, interpretations, decisions, and actions—even those based on standardized forms of assessment—are always situated in complex and partially unique learning environments, and judgments about their validity must be similarly situated. A robust validity theory and the research agenda through which it evolves must be able to consider the situated nature of IDAs.

The theoretical resources we have provided are intended to support formal inquiry, when the validity of an IDA is both explicit and documented; informal inquiry, when the validity of an IDA is explicitly considered; routine practice, when IDAs are enacted without explicit consideration of their validity (which draws attention to how
the environment is resourced); and, of course, research into the practice of assessment in different learning environments. We have distinguished between routine inquiry, as it might be enacted within a focal learning environment (such as a classroom) and as it might be enacted in an external learning environment that is responsible to or for the focal learning environment (whether informally, as with a teacher study group, or formally, as with the accountability practices of school- or district-level administrators). We have also considered questions that might be asked when evidence crosses boundaries between activity systems with and without people to recontextualize it.

We have drawn on three distinct theoretical discourses in conceptualizing validity theory for these purposes: educational measurement, hermeneutics, and sociocultural studies. These theoretical resources allow us to address (1) the soundness of particular IDAs about learning, (2) the resources of the activity systems in which they are developed and used, and (3) the opportunities for learning which different activity systems are providing their learners (students and professional educators). Within and across these theoretical discourses we see complementarities in serving the goals we have described, common understandings, and constructive disjunctions that illuminate taken-for-granted categories of thought and action for critical reflection (Bourdieu, 1991).

Educational measurement provides resources that are most directly useful when the goal is to develop a common validity argument, based on common sources of evidence and analyses, to support IDAs that are presumptively relevant across individuals and contexts. Hermeneutics provides resources that are crucial when unique, situated, and dynamic IDAs are developed based on multiple varied sources of evidence (including standardized assessments). Sociocultural theories provide resources to assist in the analysis of complex activity systems, communities of practice, or learning environments. They suggest evidence that might be gathered in response to questions about learning and questions about how assessment (or any other aspect) functions interactively in a particular learning environment. They help illuminate the dialectical relationship between social structures (e.g., assessment and accountability systems) and local practices (e.g., interactions among principals, teachers, and students or among school board members or legislators)—how each constructs, shapes, and/or challenges the other. Sociocultural studies provide a conception of educational assessment as inquiry (Delandshere, 2002) that focuses not just on what learners are learning, but also on how and why (Engeström, 2001), and that incorporates questions about what the inquirers—teachers and other education professionals—are themselves learning about supporting students’ and one another’s learning.

Common understandings about developing and evaluating sounder IDAs across these discourses include the importance of examining the processes through which IDAs are developed, seeking challenges to developing understandings in alternative perspectives and practices, developing and studying concrete examples to both illustrate and extend current understandings, and examining the consequences of our decisions and actions as distinct from our intentions. Instructive disjunctions we have highlighted include different stances that inquirers might take (different relationships
they might enact) with the people about whom they are inquiring and different perspectives on the appropriate role of methods or a priori decisions (about criteria, evidence, etc.) in inquiry. Although these discourses offer different conceptual tools—practices and perspectives—to inquirers, the alternatives, even when disjunctive, can be engaged dialectically to provide productive challenges to developing IDAs and conceptions of validity.

This constellation of theoretical perspectives relevant to validity theory holds important implications for actors who work in different communities of practice. Although we suggest sample implications here, we hope our arguments invite others who work in or with particular communities of practice to explore the implications in depth and to further develop and challenge the preliminary theoretical perspectives we have offered. For those who develop standards and curriculum frameworks intended to guide teaching, learning, and assessment, it is, perhaps, most important to provide rich descriptions and concrete examples that illustrate a full range of intended learning the standards or curriculum frameworks are intended to support. Furthermore, as an NAE panel suggested, providing access to multiple "coherent, professionally credible" standards discourages the assumption that there is "one best way to define and structure knowledge" (p. 24).

For developers of standardized assessments and those who mandate their use, it will be important to understand local users' needs and to provide them with information that will support them in making sound IDAs in their local contexts. This will include a well-specified content domain illustrated with concrete examples and, in the case where tests are intended to address standards or curriculum frameworks, explicit acknowledgment of what is not tested. Access to empirical studies that support and challenge these understandings of the tested domain will be important as well. This will allow users to consider what the mandated assessment makes visible, what is left in the background, and what the consequences of those choices are.

For teachers and those who support their preparation and development, it will be important to engage routinely in the kind of evidence-based inquiry described and to develop the norms and routines of professional practice that support critical, collaborative inquiry, and learning. For administrators, policy makers, and others who design educational environments, it will be important to consider as well how learning environments, including their own, are resourced and what opportunities they present for professional learning.

For researchers who are interested in assessment practice, there is important work to be done in studying the use of evidence by professionals working in particular contexts at all levels of the education system. What evidence do the education professionals working in these different learning environments need to know that students are learning and experiencing adequate opportunities to learn? What evidence do they need to know that the education professionals to/for whom they are responsible have the opportunities to learn and the resources necessary to support students' learning? How are the understandings and actions of education professionals shaped (enabled and constrained) by assessments and the routines that surround them? In short, the
same types of questions that we have suggested should be asked of students’ learning opportunities may equally well be asked of the learning opportunities available to educators. These will include questions about how assessment practices position them with varying authority, agency, and accountability (Greeno, 2002) and about the potential consequences of different choices. It is important to note that in raising these questions about analyzing learning environments, we have not directly addressed the question of the vision of learning that is guiding the learning environment. Our goals have been to raise questions that illuminate the “learning curriculum” (Lave & Wenger, 1991) and how the environment is functioning to support it and to highlight the critical importance of access to alternatives that illuminate the values and understandings that “we” in a given community take for granted.

The theoretical perspectives we have presented here suggest the importance of ongoing case study work to help in understanding how the always partially unique elements of any activity system interact to shape what is learned. To that end, it will be important to expand conceptions of generalizability (e.g., NRC, 2001), which support the application of generalized propositions to concrete situations, as educational measurement does, to incorporate the role that individual cases can play in developing useful knowledge. As we have argued (with theorists from each of the discourses we cite), general knowledge claims, when put to work, are always put to work in particular contexts, where evidence of how they shape and are shaped by local practice must be routinely considered. Expertise in complex professional domains does not develop only, or even primarily, through the acquisition of abstract concepts that can then be routinely applied (Beach, 1999; Bransford & Schwartz, 1999; NAE, 2005); 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. Case studies not only illustrate general principles, but they also contribute to theory by expanding our experience and our ability to raise better questions about the next case.

A central principle of our approach to validity is that we learn by seeking out perspectives, practices, and social contexts that that are different from our own; such encounters make us aware of the categories of thought and actions we take for granted; allow us to imagine how things might be otherwise; and encourage us to reconsider our perspectives and practices in light of this knowledge. It is an approach that is intended as both critical and generative. We believe there is rich opportunity for dialogue, collaboration, and mutual learning across the discourses on which we’ve drawn. We hope that practitioners of other research discourses will find ways to bring their experience to bear in developing validity theory for the practice of educational assessment. It is through such encounters that we can “learn new forms of activity which are not yet there” (Engestrom, 2001, p. 138).

ACKNOWLEDGMENTS

One of the joys of working in the multidisciplinary field of educational research, and one of its greatest strengths, is the opportunity it provides to learn from colleagues...
who hold different perspectives. I (Pamela) have benefited from collaboration with scholars in each of the three research discourses on which we draw in this document. As we write, I am simultaneously working on a chapter with Jim Greeno on the implications of sociocultural theory for the practice assessment and opportunity to learn. The ideas in this chapter have been influenced in multiple ways by my conversations with Jim and the other members of Spencer’s “Idea of Testing” Project: King Beach, Jim Gee, Carol Lee, Ed Haertel, Bob Mislevy, Bud Mehan, Fritz Mosher, Diana Pullin, and Lauren Young. Although the chapters serve somewhat different purposes, readers will find overlaps between them. Our section on hermeneutics is similarly influenced by my long-term collaboration with Aaron Schutz; readers will be able to trace many of the ideas presented here to Aaron’s and my still-growing list of jointly authored publications. Jane Cogsshall, Ray Pecheone, and Mark Wilson have contributed substantively to our work reported on the assessment of teaching. We are grateful to Judith Green, Jim Greeno, Allan Luke, Bob Mislevy, Annemarie Palincsar, Denis Phillips, and Aaron Schutz who generously provided comments on earlier drafts of this chapter. The Spencer Foundation has supported much of the work presented here, through an NAE/Spencer Fellowship, a major grant, and funds for the “Idea of Testing” conference series. We are deeply grateful for their support.

NOTES

1 Although we distinguish among different theories of validity—different approaches to developing and evaluating interpretations, decisions, and actions—we also use the term “validity theory” to describe the conceptualization of such theories. Furthermore, we note that not all theorists on whose work we draw use the term “validity”; it is our description of the term, rather than the term itself, that guided our selection of theoretical resources.

2 Our focus in this chapter is on education professionals and the communities of practice in which they work. We have not addressed equally important issues involving students and their parents/guardians as users of assessment.

3 Readers should be cautioned that categories such as these always underrepresent or misrepresent the complexity of the issues involved. Within naturalist and interpretive conceptions of social science are diverse traditions, some of which represent distinct perspectives. Although the naturalist/interpretive cut foregrounds a particular set of issues—a particular set of commonalities within and differences across traditions—a different category scheme would rearrange allegiances among traditions.

4 We will follow, initially, the Standards for Educational and Psychological Testing’s (American Education Research Association [AERA], APA, National Council on Measurement in Education [NCME], 1999) distinction between tests and assessments. The authors of the Standards define a test as: “an evaluative device or procedure in which a sample of an examinee’s behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process” (p. 3). They characterize “assessment” as a broader term, “commonly referring to a process that integrates test information with information from other sources (e.g., information from the individual’s social, educational, employment, or psychological history)” (p. 3). Our definition of assessment, however, is broader still, and refers to process of inquiry that integrates multiple sources of evidence, whether or not test based, to support an interpretation, decision, or action.

5 In understanding the potential relevance of validity theories in educational measurement for education professionals, it is important not to conflate standardization with multiple choice
tests. Essentially, standardization refers to the aspects of an assessment that are common across individuals and contexts. Although standardized assessments include multiple-choice tests, they can also include complex performance assessments, such as multimedia portfolios or observations systems, where features such as guidelines, criteria, and procedures for combining evidence are standardized.

A chapter by Phillips (in press) productively addresses many of the same issues our RRE chapter addresses.

The word "scores" is used generically to refer to both numbers and predetermined categories (such as diagnoses) assigned to individuals' performances.


Both the Testing Standards and the *Educational Measurement* volume contain multiple chapters about aspects of test development and evaluation (such as sections on test development, reliability and errors of measurement, scaling and equating, and fairness or absence of bias), which are relevant to the overall judgment that validity entails.

Earlier editions of the Standards framed the purpose somewhat differently. For instance, "the essential principle underlying this document is that a test manual should carry information sufficient to enable any qualified user to make sound judgments regarding the usefulness and interpretation of the test" (1966, p. 2).


He cautioned that general concepts such as these must be specified in terms of observable acts or behaviors of persons. Cureton noted that that some criteria, frequently criteria involved in educational objectives, may not be directly observable, because they "refer to acts which will occur (or fail to occur) long after the end of formal schooling" (p. 653) and they are stated in terms of generalities that make it hard to specify observable acts. It becomes necessary, therefore, to define more immediate objectives and intermediate criteria. However, he cautioned: "All too often the immediate objectives are derived by backward reasoning from traditional elements of the curriculum: . . . If the arbitrary immediate objectives lack ultimate relevance, such tests retard educational progress instead of stimulating it" (p. 654).

Actually, two versions were published, one on "Technical Recommendations for Psychological Tests and Diagnostic Techniques" (APA, 1954) and one on "Technical Recommendations for Achievement Tests" (AERA and NCMUE, 1955), although the conception of validity was essentially the same in both documents, so we focus on the 1954 standards.

The 1954 edition had listed predictive validity and concurrent validity as two distinct types of validity distinguished by whether the criterion measure was administered at the same time as the test in question or at some future time.

Messick (1989) depicted his conception of validity by crossing two dimensions or "facets" of validity inquiry. One facet, focusing on the function or outcome of testing, distinguishes between interpretation and use. The other facet, focusing on the justification for testing, distinguishes between appraisal of evidence and appraisal of consequence.

<table>
<thead>
<tr>
<th>Test Interpretation</th>
<th>Test Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidential basis</td>
<td>Construct validity</td>
</tr>
<tr>
<td></td>
<td>Construct validity + relevance/utility</td>
</tr>
<tr>
<td>Consequential basis</td>
<td>Value implications</td>
</tr>
<tr>
<td></td>
<td>Social consequences</td>
</tr>
</tbody>
</table>

Messick noted that this is a progressive matrix, with construct validity appearing in every cell. This highlights construct validity as the foundation or "integrating force" for validity inquiry (p. 20).
The example we used to illustrate these concepts is not taken directly from Messick’s text.

Language of title is taken from Kane (1992).

In 1992, he summarized categories of inferences in separate section of his article—and although not separately named as of general use in 2006, the same inferences appear repeatedly in different interpretative arguments: observation (or scoring), generalization, extrapolation, theory-based inferences, decisions, and technical inferences.

Mislevy (2006; Mislevy, Almond, & Steinberg, 2003), who, like Kane, draws on Toulmin’s (1958) argument analysis provides an alternative set of conceptual resources for validity theory within what we are calling validity inquiry as practical argument. We should note that both Kane and Mislevy argue that their general analysis of a validity argument, based in Toulmin’s framework, is relevant to unique interpretations based on nonstandardized assessments as in a conversation between a teacher and a student. Kane characterizes Toulmin’s general model for evaluating informal arguments as containing six elements. (1) a claim or conclusion, (2) data, (3) a warrant specifying the rule for going from the data to the claim, (4) backing or evidence to justify the warrant, (5) a qualifier that indicates the strength of the claim, and (6) conditions of rebuttal indicating the circumstances under which the warrant would not apply. Each inference in the overall interpretive argument becomes a claim that must be warranted. Mislevy’s characterization varies slightly. Readers will see both themes and variations between these conceptions and those we draw from hermeneutics below.

The senior author was a member of the committee that drafted the 1999 Testing Standards.

Although the standards are somewhat more specific about evidence needed with respect to particular technical issues (reliability and errors of measurement, scaling and equating), they are less direct about validity evidence more directly related to score meaning. Of the 24 standards in the Validity Chapter, only a few are framed as general obligations for all test users; others are conditioned on the type of evidence the developer may have deemed necessary. For example, “When the validation rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described . . .” (p. 18).

Of course, the import of this statement depends on your conception of validity.

She cited two recent textbooks (Stiggins, 2001a, and Taylor & Nolen, 2005) and a recent issue of Educational Measurement: Issues and Practice edited by Susan Brookhart (Brookhart, 2003) as providing theory and practice that begin to address teachers’ needs more directly.

Of course, different policies of test use, as distinct from educational measurement validity theory, give differential attention to considering whether an individual case fits within the intended interpretation. For advice about appropriate uses of standardized tests, see AERA et al., 1999; Koretz & Hamilton, 2006; Herman & Haertel, 2005.

Mislevy et al., 2003 accomplish a similar purpose and provide an extended example from the domain of second language testing.

They cited, as well, purposes for which students and parents used assessment, including for students “deciding whether they were capable of learning” and whether it was worth trying, and for parents, “when to reward and punish, how to allocate family resources, and whether to seek additional help” (p. 9).

The report contains multiple brief examples in every chapter to illustrate the committee’s recommendations. Examples of assessments that were intended as at least partially consistent with the vision of assessment found in Knowing What Students Know can be found in two volumes edited by committee member Mark Wilson (2004, Wilson & Bertenthal, 2006), and a Web site developed by committee co-chair James Pellegrino is http://aim.pschi.uic.edu/.

Subsequent articles by committee member Bob Mislevy and colleagues (e.g., Mislevyet al., 2003) offer extended advice for building a validity argument, within the practical argument framework, that can support assessment involving complex forms of evidence and dynamic interpretations and, Mislevy argues, multiple visions of learning.

Many, but not all, sociocultural approaches to learning and assessment trace their roots in part to the work of Vygotsky and his arguments for the role of social interaction in learning and
the notion that social interaction precedes mental representations: “Every function in the child’s cultural development appears twice: first, on the social level, and later, on the individual level; first, between people (interspsychological) and then inside the child (intrapsychological). This applies equally to voluntary attention, to logical memory, and to the formation of concepts. All the higher functions originate as actual relationships between individuals” (p. 57). Wertsch (1995) further notes that Vygotsky did not, himself, use the word, sociocultural.

20 Examples of more standardized approaches to assessment have been proffered as consistent with a sociocultural approach to assessment. With dynamic assessment (e.g., Magnusson, Templin, & Boyle, 1997), the assessment is designed to scaffold the learner’s performance, providing whatever social support is necessary to complete the task successfully. The performance is evaluated of the extent of social support the learner needs to be successful. Hickey and Zuiker (2003) provide multiple examples of group interactions, typically video recorded, that are evaluated for various criteria of participation. They argue that these studies of interaction might be used, dialectically, with more traditional approaches to assessment and with in-depth ethnographic research into a learning environment. Mislevy (forthcoming), drawing on his Evidence Centered Design approach to large-scale assessment consistent with sociocultural perspective, points to the portfolio assessments used in AP Studio Art (see also Myford & Mislevy, 1995). The AP Studio Art examples allows substantial flexibility for students to present a coherent series of pieces that represents their own areas and themes of specialization, which are then centrally scored with a standardized rubric. As such, he argues, it takes context and students into account.

21 As Shepard (2006) noted, “assessment for learning” was a term coined by the Assessment Reform Group of the British Educational Research Association.

22 This description is excerpted from an extended case study based on Lampert (2001) in Moss and Greeno (forthcoming).

23 Numerous scholars in educational measurement and assessment have pointed to interpretive research traditions for supporting the validity interpretations based on nonstandardized forms of evidence, especially in the context of classroom assessment (Delandshere, 2002; Gipps, 1999, 2002; Kane, 2006; Mislevy, 2006, in press; Mosset et al., 2003; Shepard, 2001, 2006).

24 The importance of concrete examples or cases of practice has been highlighted by theorists from each of the traditions on which we draw (e.g., Shepard, 1993; Gadamer, 1994; Lave & Wenger, 1991).

25 This section draws on Moss, 2005; Moss and Schutz, 2001; Moss, Cogshall, and Schutz, 2006.

26 It is beyond the scope of this review to draw the connections between hermeneutics and other characterization of validity within “interpretive” or “qualitative” research. Interested readers might consider, for instance, the chapters on qualitative research in the two recent editions of the Handbook for Research on Teaching (Erickson, 1986; Lather 2004) to see the extent to which it is possible to develop an interpretation of the differences in these approaches in the language developed here.

27 Such interpretations can incorporate results from standardized tests and benefit from the research that presumptively supports them, contextualized, as educational measurement validity theory suggests, in locally relevant evidence.

28 As characterized by Ormiston and Schriff (1990), the correct meaning was typically conceptualized as the original intent of the author (Schleiermacher) or as the events or objects experienced by the author (Dilthey). Betti (1990) and Hirsch (1976) provide more recent arguments for this approach to hermeneutics. (Note copyright dates do not reflect chronology.)

29 In his earlier work, Habermas (1984) called this the “ideal speech situation,” although he has since abandoned this label because of the misinterpretations it provoked (1993, pp. 163–164).

30 The debate between Gadamer (1990) and Habermas (1990a) about the ideal speech situation and the role of explanatory discourse is instructive and offers interpreters alternative stances (see Ormiston & Schriff, 1990). Gadamer worried that Habermas’s emphasis on rationally
motivated consensus is a "fantastic overestimation of reason by comparison to the affections that motivate the human mind" (1994, p. 567). Gadamer also worried that the move to explanatory discourse or depth hermeneutics, as Habermas framed it, signals a failure in conversation—instead of talking with the other person, we are talking about them, and we can no longer "be reached" by the text. That this may happen does not absolve us of the initial obligation to search for coherence and for what we can learn from what they have said. Although both Habermas and Gadamer believed that unconscious prejudices distort understanding and that they must be somehow "provoked" (Gadamer, 1987, p. 137) and made explicit, Gadamer believed that this could be accomplished through the encounter between an interpreter and a text, as long as the interpreter approaches the text with hermeneutic attitude—the willingness to acknowledge the coherence and truth in the text, and thereby risk his or her own prejudices. For additional discussions of philosophical and depth hermeneutics see Flyvbjerg (2001), Ricoeur (1990), and Thompson (1990).

40 Our descriptions of sociocultural theories overlap substantially with those in Moss and Greeno (forthcoming).

41 Seminal theoretical resources can also be found in the work of Bruner (1990), Cole (1996), Rogoff (1990), and of course, in the historically relevant texts of Leontiev (1978), Luria (1979), and Vygotsky (1978, 1986). Accessible introductions to newcomers can be found in Lave and Wenger (1991); Wertsch, Del Rio, and Alvarez (1995); and Wells and Claxton, (2002).

42 Figure taken from Gee, forthcoming.

43 As Mislevy (personal communication) noted, measurement theorists might well raise questions about the quality of diagnostic information at the subdimension level on large-scale standardized tests, in part because of questions about the reliability of subdimensions and in part because of questions about connection to the local curriculum.


45 A forthcoming NSSE Yearbook on “Evidence and Decision Making” (Moss, in press) and a separate volume on “Opportunity to Learn” (Moss, Pullin, Gee, Haertel, and Young, forthcoming) provide additional illustrations that focus on evidence use at different levels of the educational system.

46 Well-designed standardized assessments and the validity inquiry that supports them can serve numerous important purposes for educational professionals working in different contexts. For instance, they can provide carefully researched examples of sound assessment practice, they can permit economies of scale that make use with larger numbers of individuals or institutions feasible, they can raise hypotheses about or identify cases for further study, they can provide a synoptic view to support decision making by administrators and policy makers who work at centralized levels of the educational system, and they can enable comparisons across individuals and contexts that are crucial to addressing questions about equity.

47 Although it is beyond the scope of this chapter to review, we note that activities involved in the development of standardized assessments frequently resemble these dynamic and integrative interpretive practices (Moss & Schutz, 1999; Mislevy, personal communication). These include the IDAs taken during development of tasks, scoring rubrics, strategies of aggregation and analysis, standard setting, and so on.

48 Conversations with Fred Erickson, Jim Gee, and Bob Mislevy have influenced my (Moss) conceptions of generalizability.
REFERENCES


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.


