The Role of Consequences in Validity Theory

Pamela A. Moss
University of Michigan

How do individuals make sense of and use the products and practices of testing in their everyday lives? What is the responsibility of the educational measurement community to take these issues into consideration in assessing what it is that we do?

> People know what they do; they frequently know why they do what they do; what they don't know is what what they do does.
>
> —Foucault, cited in Dreyfus and Rabinow, 1983

The role of consequences in validity theory remains a matter of considerable debate. While virtually all validity theorists acknowledge the value of research into the consequences of test use, there are substantial differences in perspective about whether and how attention to consequences should be incorporated into the concept of validity. Contrasting the perspectives of Cronbach (1988, 1989), Messick (1989, 1994, 1996), Shepard (1993, 1997), and Wiley (1991), for instance, highlights differences on the following range of issues.

- Should consequences be considered an aspect of validity at all? If yes, how should the concepts be related?
- Are inappropriate consequences relevant to validity only if they can be traced to a source of construct underrepresentation or construct irrelevant variance?
- Or could negative consequences associated with an otherwise valid interpretation call the validity of the test use into question?
- If yes, does that mean we have expanded the focus of validity from the focus on an assessment-based interpretation to a focus on the larger system of which the assessment is a part?
- On what grounds should these decisions be made? Epistemological? Ethical? Practical? Political?

Messick (1996), for instance, argues that "the primary measurement concern with respect to adverse consequences is that any negative impact on individuals or groups should not derive from any source of test invalidity such as construct underrepresentation or construct irrelevant variance" (p. 13). He goes on to note, however, that "if found, one should monitor the situation to see how short-term it is likely to be and what resources are needed to redress the imbalance" (p. 13). This suggests that adverse consequences undermine the validity of an assessment only if they can be traced to a problem with the fit between the test and the construct. [Here, I should note that Messick, who is frequently cited in conjunction with the concept of *consequential validity*, actually eschews that term (1996). He refers, instead, to the consequential *aspect* (1994, 1996) or *basis* (1989) of validity which highlights validity as a unitary concept for which evidence about consequences is only one part.]

In apparent contrast to Messick, Cronbach (1988) suggests that a test interpretation that "honestly reports facts" (p. 5) is open to validity challenges whenever adverse consequences arise. "Tests that impinge on the rights and life chances of individuals are inherently disputable" (p. 6). Elsewhere, Cronbach suggests questions of social consequence expand the focus of validity to include the whole system of which the test is a part (Cronbach, 1980, p. 101, 103). For Cronbach, it seems, adverse social consequences, in and of themselves, may call the validity of a test use into question. Wiley (1991), taking yet another position, argues that attending to such adverse consequences (or what he terms *use errors*, p. 88), while socially important, "would needlessly complicate the conception and definition of test validity" (p. 88). Shepard (1997) responds to such arguments by pointing out that the consequential aspect of validity is hardly a new concept: "Consequences," she argues, "are a logical part of the evaluation of test use, which has been an accepted focus of validity for several decades" (p. 5). She suggests further (1993) that potential adverse consequences or other unintended effects are simply rival hypotheses to the express purpose of testing.

Each of these theorists articulates a different perspective on the relationship between validity and consequences and builds an argument on somewhat different grounds. These are differences that could lead to different conclusions about the degree of *validity* associated with a given interpretation or use of a test. And so, the question of whether to incorporate consideration of consequences into the definition of validity is not just an interesting philosophical question; it can be seen to have real ethical, political, and economic consequences. There are no easy answers

Pamela A. Moss is an Associate Professor in the School of Education, University of Michigan, 610 East University, Ann Arbor, MI 48109-1259. Her specializations are educational measurement and evaluation.

about the responsibilities of validity researchers with respect to evidence about consequences. Allocating resources to the study of consequences takes them away from something else that may be equally or more valuable to the educational community. That's why it's so important to have the kind of dia-logue that the articles in this issue engage.

Overview

In the articles that follow, my colleagues will present systematic and specific suggestions for structuring the activities and responsibilities for research into the consequences of testing. For this article, I have two interrelated purposes that I hope will complement that advice. The first purpose is to provide an argument for incorporating consideration of consequences into validity theory that is grounded in the reflexive nature of social knowledge. It focuses on the ways in which the interpretations of social scientists (including test developers, users, and evaluators) can be and often are reinterpreted and integrated into the lives of the subjects theydescribe—and the ways in which social reality can be transformed in the process (Thompson, 1990; see also Bourdieu, 1990; Foucault, 1977; Hoy, 1994; Luke, 1995). To the extent that the practices in which we engage change the social reality we study, the study of consequences becomes an essential aspect of validity even for those who choose to limit the scope of validity to a test-based interpretation. The second purpose of this article is to argue for the importance of considering a set of questions and type of evidence we don't typically pursue concerning the consequences of testing—evidence based on the actual discourse that surrounds the products and practices of testing. Such evidence allows us to question both how individuals make sense of the information they receive and how this might impact the way they understand themselves and others (Gee, 1996; Luke, 1995; Mehan, 1993). Because these arguments—about the importance of evidence based on discourse in context and the impact of testing practices on social reality—are mutually supportive, the sections that follow will interweave theoretical argument with three concrete illustrations of how and why we might study the effects of the products and practices of testing.

Clearly, this perspective on the consequences of testing spills over the consensual boundaries of validity—which encompass the evaluation of specific interpretations and (for many theorists) uses of tests scores—to include evaluation of the consequences of testing more generally. This pursues a path to which Messick pointed in his 1989 chapter on validity: "We will underscore the continuing need for validation practice to address the realities of testing consequences, including the often subtle systemic effects of recurrent or regularized testing on institutional or social functioning," (Messick, 1989, p. 18). While I believe this argument has some practical implications for those who develop and use tests (to which I will point later on), the bulk of the responsibility for this sort of theoretical and empirical research must fall on the measurement profession at large. It is for this larger, long-term research agenda about the consequences of testing that I argue.

Studying the Impact of the Products and Practices of Testing

Theoretical Explication 1: The Major Issues

If we want to better understand the consequences of test use, we need to understand how individuals make sense of and use the products and practices of testing in their everyday lives. Here, I refer to all aspects of testing as experienced in the local context—including, tasks, administration, scoring, interpreting, and using test scores—although in the context of this article, I'll focus primarily on the messages contained in score reports. While our validity research typically focuses on establishing the validity of fixed interpretations of test scores, the meaning of these messages in local contexts is not a fixed property of the message itself. Rather, it depends on how the individuals draw on the resources available to them in their particular sociohistorical circumstances to understand the messages they receive (Thompson, 1990). The consequence of the dissemination of these messages depends, in turn, on how individuals incorporate these messages into their daily lives—how these messages affect the way they understand themselves and others (Thompson, 1990). The structure of the message itself—not just in terms of its intended meaning but also in terms of its (not always intended) implications regarding the role of readers and the nature of knowledge—contributes to, but does not determine, its effects. This set of research questions about the meaning and effects of test-based interpretations in local contexts is not something we can address by asking individuals what they think in the structured and standardized ways to which we are accustomed. By doing that, we lose our ability to understand the ways in which they might represent themselves, without our concepts and categories, or the way in which our categories may simply mean something different to them. Rather, we need to study the actual discourse and actions that occur around products and practices of testing.

Illustration 1: A Definition of Validity

To illustrate these points, I will characterize a situation where we—those of us in the measurement profession—are subject to the consequences of a particular widely disseminated interpretation. And then, using that as an analogy, I'll turn to the issue at hand: the possible consequences of test-based interpretations to the individuals who make meaning of them in the contexts of their daily lives.

Here are a few sentences excerpted from the first paragraph of the validity chapter in the 1985 *Standards for Educational and Psychological Testing*. While reading them over, consider what the text implies about the positions of readers and writers and about the nature of knowledge about validity:

> Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores . . . . Validity . . . is a unitary concept. Although . . . evidence may be accumulated in many ways, validity

> always refers to the degree to which that evidence supports the inferences that are made from the test scores. (APA, AERA, & NCME, 1985, p. 9)

The first thing to note is that there is no use of the first person plural. There is no reference to the writers behind the definition. Second, the definition of validity is presented nonproblematically, as a given. There is no reference to or consideration of the evolving history of the concept or the competing alternatives that appeared in the literature of the time. There is nothing in the passage that invites readers to assess the definition for themselves. What, we might ask, are the consequences of representing validity in that way to individuals within our profession? This passage, I could argue, constructs readers as passive recipients of a pre-existing concept and hides the power exercised by the writers over the representation of the concept and the authority to question it. Over time, perhaps this passage and the passages like it that we encounter may influence the way we understand and enact our roles as measurement professionals.

This conclusion about the potential impact of the passage could be dismissed as speculation, but we could think about the kind of evidence we might collect to evaluate its validity. For example, I'd want to listen to the conversations in graduate classrooms and test development conference rooms where the document is used. I'd want to know who knows about the document but dismisses or rejects it and why. I'd want to know who never heard of the document and how and why they represent validity as they do. I could turn to the literature on validity theory, and I might notice, for instance, that a very small proportion of the members of our profession write articles that attempt to criticize, elaborate, or extend validity theory. To what extent has the practice of defining validity in the manner of the 1985 *Standards* influenced that outcome? We'll never have a definitive answer, because the potential influences are so complex and inter-related. However, that should not detract from the importance of raising and pursuing questions about the consequences of representing validity in this way.

In making this argument, I don't mean to imply this representation of validity is necessarily bad; in fact I have, in comfortable conscience, written my own fair share of similar sentences. Inviting readers to develop their own perspectives about the nature of validity could result in fostering an "anything-goes" mentality and lead to considerably worse practices in validity research. My point is simply that we should work to make ourselves aware of the potential consequences of different choices and consider them explicitly in deciding on the appropriateness of the practices in which we engage.

Theoretical Explication 2: The Dialectical Relationship Between Social Reality and Its Representation

The social theorists I read offer extended philosophical arguments and empirically based illustrations of the dialectical relationship between social reality and our representation of it (Bourdieu, 1990; Foucault, 1977, 1980; Taylor, 1987). As Thompson (1990) states, the interpretations that social scientists construct can be, and often are, reinterpreted and integrated into the lives of the subjects they describe. And in that process, "the domain may itself be transformed" (Thompson, 1990, p. 276). This suggests how important it is to understand the extent to which the test-based interpretations become "part of the taken-for-granted definitions and categories by which members of communities define themselves and others" (Luke, 1995, p. 9) and to consider the political and ideological consequences to individuals of being represented and representing themselves in that way. As Luke (1995) suggests, "the repertoire of representations, practices, and positions made available to students in turn has identifiable material consequences for those students" (p. 21). Over time, "these devices operate by building up a version of the social and natural world and, at the same time, constructing social relationships of power" (p. 17). Similarly, Wacquant, paraphrasing Bourdieu (1992), argues: "If we grant that symbolic systems are social products that contribute to making the world, that they do not simply mirror social relations but help constitute them, then one can, within limits, transform the world by transforming its representation" (p. 14). In this observation lies both the positive potential of and the danger of testing practices. That is why it is so crucial to understand the consequences of testing at the level of discourse in context.

Illustration 2: Individual Score Reports

Turning again to a concrete example, consider the following excerpts from two score reports that were intended for parents or guardians and their children. One is from a state mandated criterion-referenced test, and the other is from a commercially available nationally normed achievement test. Both focus on mathematics achievement scores. The criterion-referenced test locates the student's overall score in mathematics in one of three categories: low, moderate, and satisfactory. Parents are told that the results are reported "in relation to a standard set by experienced educators." On this particular report, the child received a score which fell in the satisfactory range. The parents are told:

> Your child understands important mathematical concepts, and can select and apply mathematical operations presented in both number sentences and word problems. Although your child may have missed some test items, this score suggests that your child is well prepared for further study.

Had the child received a score that fell within the low range, the parents would have been directed to read the following information:

> Your child is not well prepared in mathematics. This could be because your child has not had the opportunity to explore all of the ideas and skills tested, OR the manner in which those concepts and skills were taught has not been meaningful for your child, OR your child has not applied the continued attention, motivation and effort needed to achieve this standard.

The norm-referenced test report in mathematics offers the following computer generated interpretation of the student's national percentile and grade equivalent scores:

> These scores provide a way of comparing your child's test performance . . . with a national norming sample of students tested in [year].
>
> Your child's . . . NATIONAL PERCENTILE SCORE for Mathematics Computation was 54 . . . . This means his/her performance on this subtest was higher than 54 percent of the 5th grade students in the norming sample . . . . The GRADE EQUIVALENT SCORES are reported in grade levels and months. This test was given in the 5th grade during the 7th month of the school year, April. This means that an average 5th grade student in the norming sample would have a score of 5.7. Your child's . . . grade equivalent score in math computation was 6.1. This means that your child scored as well as the average student taking this same test in October of the Sixth Grade.
>
> It is important to remember that grade equivalent scores do not say anything about what grade your child should be placed in. These scores only reflect your child's performance on 5th grade material; they do not tell us how your child would perform on material from another grade level.

Both reports give information on the child's performance on subsets of items, some of which are as small as five, listing the number correct divided by the total numbers of items. The norm-referenced test also provides an indicator of mastery for each subset of items (objective) along with a key that defines mastery and partial mastery with cut scores at 75% and 50%, respectively.

Now, while the norm-referenced test offers an interpretation that one might argue stays closer to the evidence likely to be underlying these scores, both score reports assert brief authoritative messages about the performance of these children. Here, again, we might ask the same sorts of questions that we asked about the definition of validity. What do these reports imply about the positions of readers and writers and the nature of knowledge provided about these children? What are their potential effects?

If we want to understand the consequences of the use of this test, we need to know what happens when and after a parent or guardian opens this report. How do they make sense of the information? What stance do they take with respect to the report? What do they say to one another, to their child? Do they accept the interpretation as given? Do they consider the potential error associated with the score? Do they consider alternative explanations for the interpretation? Do they ignore or dismiss it? Do they share the information with others? And if so, how? Does their interpretation of their child's capabilities change? Does this information influence the way they interact with their child in the future about school, or homework, or future opportunities? Do the answers to these questions vary depending on the educational, economic, or sociocultural background of the parents or guardians?

To reiterate, my point is not simply whether they are understanding the report in the way that its authors intended—although that's an important question. Equally important are questions about the stances this report invites them to assume and the ways in which they incorporate this information into their daily lives.

Theoretical Explication 3: A Conceptual Framework for Studying Discourse in Context

To provide some guidance for studying these issues, I've borrowed from the work of critical social theorist, John Thompson (1990), to suggest a set of questions for studying the way in which standardized test results are interpreted and used in local contexts and the way in which these contexts may, in turn, be altered in the process. Thompson's work focuses on mass communication. The principle characteristics of mass communication that Thompson describes share characteristics in common with the context of large scale assessment. These include the separation between the contexts of the producers and receivers and the one-way only flow of communication which requires the production of an extended message in the absence of any response from receivers. Thompson suggests an approach which focuses on three domains of analysis: the production and transmission of the message, the structure of the message itself, and the reception and appropriation of the message. (By appropriation, he refers to the way in which receivers reinterpret and use the message in their own terms.) Given the issues addressed in this article, I draw primarily on the domains that deal with the processes of reception and appropriation and to a lesser extent the message itself (although careful study of the domain of production would also support critical reflection about the theories and practices of educational measurement). Central to Thompson's and other critical theorists' (e.g., Kogler, 1996) perspectives is the role of dialogue between researcher and researched. For them, the validity of the interpretations rests in part on a respectful attention to the self-understanding of those researched and in part on their critical response to the interpretations produced.

The analysis Thompson (1990) suggests for studying the reception of mass-mediated messages in local context includes a combination of social-historical analysis and ethnographic research.

> By means of social-historical analysis, we can examine the specific circumstances and the socially differentiated conditions within which media messages are received by particular individuals. The specific circumstances: in what contexts, with what company, and what degree of attention, consistency and commentary, do individuals read books, watch television, listen to music, etc.? The socially differentiated conditions: in what ways does the reception of media messages vary according to considerations such as class, gender, age, ethnic background and the geographical location of the recipient. Such social-historical analyses can be conjoined with a more interpretive form of inquiry in which we seek to elucidate how particular individuals, situated in specific circumstances, make sense of media messages and incorporate them into their daily lives. This interpretation of the everyday understanding of media messages may help to highlight the rules and assumptions which recipients bring

> to bear upon media messages, and by means of which they understand these meanings in the way that they do. It may also help to highlight the consequences which media messages have for the individuals who receive them, including the consequences for the relations of power in which these individuals are enmeshed. (pp. 305–306)

Illustration 3: A School-Level Report

To illustrate this approach,Thompson (1990) draws on Radway's study of readers of romance fiction. In this section, I adapt and extend Thompson's example to a more relevant circumstance: the reception and appropriation of a school-level test report by members of a public school faculty. The next six paragraphs list categories of issues/evidence closely paraphrased from Thompson's framework and related questions (appropriated from his example) that might be raised in understanding the effective meaning and consequences of test reports.

Following Thompson's advice (1990, pp. 313–318), if we wanted to understand the effects of the dissemination of a school-level test report on members of a faculty, we might collect evidence about:

1. *The meaning of the message as interpreted by the recipients,* including the specific ways in which they attend to (or ignore) the message and the stance they take with respect to the message:

- How do members of the faculty interpret the information received in the report? What sense do they make of it?
- To what extent is this consistent with what the test publisher intends? How do they evaluate the report? (Do they endorse the report? Do they reject it?)
- How do they attend to the report? (Do they study it carefully and care about its contents? Do they give it a passing glance and turn their attention to other things? Do they ignore it altogether? Are they angry or elated or resigned? Are they critical of the contents? Are they confused?)

2. *The acquired knowledge that individuals use to understand the message:*

- What skills or technical capabilities are required to understand the report, and to what extent do recipients have access to those resources?
- What background knowledge do they draw on in interpreting the report? (Knowledge about educational testing? About the specific content and objectives of the test? About alternative theories of knowledge or pedagogy in the subject matter tested? About the curriculum in the school or as enacted in classrooms? About the resources available to the school? About the interests of the policymakers who implemented the test? About the seriousness with which the students in their classrooms treated the test?)

3. *The actual circumstances in which the message is received:*

- How is the report transmitted? As a print document? Via computer? Interpreted orally by someone else? A televised press briefing?
- When they receive the report, are they alone or in the company of others? If in the company of others, under what circumstances? (Is it a formal faculty meeting, a one-on-one meeting with the principal, an informal gathering in the faculty room or near the mailbox, in the presence of their students?)
- At what time, according to what schedule, and in what place is the report received? How is the space configured? (Rows of chairs facing a lectern? Across the desk in the principal's office? Furniture informally gathered in the faculty lounge?)

4. *The sociohistorical characteristics of the contexts in which the message is received,* including the social institutions within which the message is received, the rules and convention which govern reception practices and related patterns of interaction, and the relations of power and the distribution of resources among individuals:

- In what institutional contexts are the messages viewed? A school? A union office? A public press briefing?
- Who decides whether or not to participate in the testing program? Who receives the report and decides when and how it is disseminated? In what ways is it appropriate for members of the faculty to respond to the report? (To reflect on the message for themselves? To accept the information without question? To consider possible action? To do what they are told in light of the information? To express an emotional reaction such as excitement or anger or chagrin?)
- What is the relationship among individuals who receive and engage in dialogue about the message? Are they peers, or is there a social hierarchy (e.g., adult and child, teacher and principal, policymaker and test developer)?
- Who gets to speak about the message, when, and to whom? What is appropriate to say?

5. *The forms of interaction and mediated quasi-interaction about the message,* including (a) interaction among individuals who received the message directly, (b) interaction involving individuals who did not receive the message directly, and (c) the virtual community of recipients who may not interact with one another directly or indirectly but who share in common the fact that they received the same messages:

- Who receives copies of the report directly, and who hears about it from others?
- What kind of interaction, if any, occurs about the report among those who heard it directly? Between those who heard it directly and others? How do those others interpret the now-mediated message?
- How is the original information communicated to others who may not have directly received the original report? (In informal conversation? Through a formal presentation? Through the news media?)
- How do those who receive the report feel toward others in the school who received the same score report? Toward others who received a score report with better or worse news? (Solidarity? Competition?)

6. *The discursive elaboration of the mediated messages,* including the ways in which the message is transformed through a process of telling and re-telling, and interpretation and criticism:

- When the report is discussed or otherwise described, what is the substance of the re-interpretation? What is the understanding reflected in the message about the message?
- How is the information subsequently used? What decisions or actions are taken based on the report? What other information in-

forms these decisions and actions?

- With what authority is the meaning of the message communicated? What stance does the person interpreting the message take with respect to the message? One of knowledgeable authority? Of critical reflection? Of confusion? Of dismissal? Of rejection (Thompson, 1990, paraphrased and appropriated from pp. 313–318)?

Taken together with a careful analysis of the structure of message in the test report itself, these aspects of discourse in context enable us to begin to understand the consequences of the interpretations disseminated in terms of how individual stakeholders make sense of them, evaluate them, and integrate them into their daily lives. How might these messages be affecting the way they understand themselves and their positions in the social hierarchy? How might the messages be affecting the way they understand others and participate in their communities? When we consider how the answers to these questions might vary across individuals and contexts that differ with respect to factors such as access to educational resources or sociocultural background, we also raise questions about the extent to which assessment practices might be enmeshed in reinforcing social inequities.

Implications for Validity Theory

This perspective on the dialectical relationship between social reality and our representation of it has implications for understanding the crucial role of evidence about consequences in validity research. I heard one NCME colleague, in arguing for the return to a more traditional view of validity, advise that we should simply stop using assessments as a policy lever to promote change. While we may be able to alter our self-conscious intentions, this will not make the effects go away. It will simply put them to work "behind our backs" (Gadamer, cited in Bleicher, 1980, p. 112). The practices in which we engage help to construct the social reality we study. While this may not be apparent from the administration of any single test, over time the effects accumulate. Foucault (1977) paints a provocative picture of how evolving practices in social science, including the use of examinations, have radically altered our conceptions of individual identity and enhanced our ability to monitor and control people's actions. Unless we work to illuminate the subtle mechanisms and outcomes of this influence, we risk both misconstruing the effective meaning of our interpretations to those who receive them and participating in the construction of a social reality that we may not intend. Returning to the quote from Foucault with which this article opens, "people know what they do; they frequently know why they do what they do; what they don't know is what what they do does" (Foucault, cited in Dreyfus and Rabinow, 1983, p. 187). As I argued in the introduction, to the extent that the practices in which we engage change the social reality we study, the study of consequences becomes an essential aspect of validity even for those who choose to limit the scope of validity to a test-based interpretation.

While this argument has some implications for validity theory as it relates to specific interpretations and uses of test scores, the import spills over this consensual focus of validity theory to encompass the general practice of testing. When the study of discourse in context informs us about the extent to which the actual interpretations of test scores are consistent with the intended meaning or about the extent to which the effects of test use are consistent with the intended purposes of testing, then the specific validity argument is implicated. However, the scope of the argument goes well beyond these test specific evaluation practices; it entails an ongoing evaluation of the dialectical relationship between the products and practices of testing, writ large, and the social reality that is recursively represented and transformed.

Implications for Practice

With respect to specific testing programs, I do think that all of us who mandate, develop, and use tests have an obligation to consider how they might be incorporated into the particular contexts in which they are implemented. While it will not be feasible for most test developers, users, or policymakers to undertake systematic research of the sort I have in mind, it is possible to anticipate particular effects, drawing on existing literature and experience, and to try to develop practices and products that enhance the positive effects while guarding against the negative ones. To the extent that case studies of the use of test-based information in local contexts are possible, validity evidence regarding specific interpretations and uses will be enhanced.

Beyond these test-specific questions, those of us in the measurement profession should consider the importance of studying the consequences of the repeated and pervasive practices of testing. Clearly, what I am proposing is intensive, highly contextualized, sustained interpretive work. While many of us may not have the resources to undertake this kind of work ourselves, we can at least (initially) seek to develop collaborations with those who do. We have colleagues in AERA who engage in and find funding for this sort of research regularly. And, we can certainly read more widely to consider the possibilities for this kind of work. This emphasis on the value of an outside perspective to illuminate what is taken for granted and thereby to provoke critical self-reflection is a theme that resonates across multiple philosophies of social science (e.g., Gadamer, 1987; Hoy, 1994; Kogler, 1996; McCarthy, 1994; Messick, 1989; see Moss, 1996, in press, for an elaborated discussion).

As we consider the benefits of this long-term research agenda, it is important that we not expect reassuring generalizations or systemic reforms that would ensure the elimination of negative consequences. Rather, the goal is to develop specific concrete examples that will enhance our understanding about the ways in which tests can and do work in local contexts and about the potential slippage between what we well-meaningly intend and what we in fact effect. If we think of it as a long-term research agenda, to elaborate our conceptual framework about the possibilities and risks associated with test use, then we will be engaged in a generative program of critical reflection likely to enhance the value of our work for those we study and serve.

Notes

This article was presented at the Annual Meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago, March 1997. I am grateful to Martin Packer for drawing my attention to Thompson (1990) as a methodological resource for critical theory.

References

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

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

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

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

Cronbach, L. J. (1980). Validity on parole: How can we go straight? *New directions for testing and measurement: Measuring achievement progress over a decade* (pp. 99–108). San Francisco: Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

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

Dreyfus, H. L., & Rabinow, P. (1983). *Michel Foucault: Beyond structuralism and hermeneutics* (2nd ed.). Chicago: University of Chicago Press.

Foucault, M. (1977). *Discipline and punish* (A. Sheridan, Trans.). New York: Vintage Books.

Foucault, M. (1980). Two lectures. In C. Gordon (Ed.), *Power/Knowledge: Selected interviews and other writings by Michel Foucault* (pp. 78–108). New York: Pantheon Books.

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

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

Hoy, D. C. (1994). Critical theory and critical history. In D. C. Hoy & T. McCarthy (Eds.), *Critical theory* (pp. 101–214). Oxford, UK: Blackwell.

Kogler, H. H. (1996). *The power of dialogue* (P. Hendrickson, Trans.). Cambridge, MA: The MIT Press.

Luke, A. (1995). Text and discourse in education: An introduction to critical discourse analysis. *Review of Research in Education, 21*, 3–48.

McCarthy, T. (1994). Philosophy and critical theory: A reprise. In D. C. Hoy & T. McCarthy (Eds.), *Critical theory* (pp. 101–214). Oxford, UK: Blackwell.

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

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

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

Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1–18). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

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

Moss, P. A. (in press). Recovering a dialectical view of rationality. *Social Indicators Research.*

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

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5–8, 13, 24.

Taylor, C. (1987). Interpretation and the sciences of man. In P. Rabinow & W. M. Sullivan (Eds.), *Interpretive social science* (pp. 33–81). Berkeley: University of California Press.

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

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in the social sciences: A volume in honor of Lee J. Cronbach* (pp. 75–107). Hillsdale, NJ: Erlbaum.