The Gordon Commission
on the Future of Assessment in Education

Postmodern Test Theory

Robert J. Mislevy
Educational Testing Service

(Reprinted with permission from ***Transitions in Work and Learning: Implications for Assessment, 1997***, by the National Academy of Sciences, Courtesy of the National Academies Press, Washington, D.C.)

Good heavens! For more than forty years I have been speaking prose without knowing it.

Molière, *Le Bourgeois Gentilhomme*

INTRODUCTION

Molière's Monsieur Jourdan was astonished to learn that he had been speaking prose all his life. I know how he felt. For years I have just been doing my job— trying to improve educational assessment by applying ideas from statistics and psychology. Come to find out, I've been advancing "neopragmatic postmodernist test theory" without ever intending to do so. This paper tries to convey some sense of what this rather unwieldy phrase means and offers some thoughts about what it implies for educational assessment, present and future. The goods news is that we can foresee some real improvements: assessments that are more open and flexible, better connected with students' learning, and more educationally useful. The bad news is that we must stop expecting drop-in-from-the-sky assessment to tell us, in 2 hours and for $10, the truth, plus or minus two standard errors.

Gary Minda's (1995) *Postmodern Legal Movements* inspired the structure of what follows. Almost every page of his book evokes parallels between issues and new directions in jurisprudence on the one hand and the debates and new developments in educational assessment on the other. Excerpts from Minda's book frame the sections of this paper. They sketch out central ideas in postmodernism—neopragmatic postmodernism, in particular—and how they are transforming the theory and practice of law. Their counterparts in assessment are discussed in turn.

Modernism and Postmodernism

This section introduces the terms "modernism," "postmodernism," and "neo- pragmatic postmodernism" as I will use them. It is necessarily incomplete, and adherents of each position will find much to disagree with.

Modernism

Modern legal theorists believe that they can discover the "right answers" or "correct interpretation" by applying a distinctive legal method based on deduction, analogy, precedent,

interpretation, social policy, institutional analysis, history, sociology, economics, and scientific method. . . . Legal moderns . . . express the intellectual and artistic quest for perfection through the process of *uncovering* and *unmasking* the secrets of the world by transcending contexts that limit human understanding. . . . Legal modernism also . . . is based on an understanding of language that assumes that words and conceptual ideas are capable of objectively capturing the meaning of events the law seeks to describe and control. (Minda, 1995:5-6)

To Plato the nature and intelligibility of the world of appearance could be accounted for only by recognizing it as an "image" of the truly intelligible structure of being itself. These "forms" are the essence of being in the world, although we experience only images or imperfect instances of this or that. He likened our condition to that of dwellers in a cave, who see shadows on cave walls but not the objects that cast them. The struggle, by means of logic and the scientific method, to infer the universe's "true" forms and to explicate their invariant relationships to experience characterizes what we may call the modern approach.

Modernism in physics, for example, can be illustrated by the prevailing belief, up through the beginning of the twentieth century, that objects exist in a fixed Euclidean space and interact in strict accordance with Newton's laws. Measurement was a matter of characterizing properties of objects such as their mass and velocity—with uncertainty to be sure but only from the imperfections of our measuring devices. The variables were the universe's; the distance between our knowledge and the truth was quantified by a standard error of measurement and shrank toward zero as we fine-tuned our models and improved our instruments.

In law the essence of modernism is the idea that "there is a 'real' world of legal system 'out there,' perfected, formed, complete and coherent, waiting to be discovered by theory" (Minda, 1995:224). The source was debated, to be sure: Dean Christopher Columbus Langdell (in 1871) maintained that careful study of cases should reveal the underlying axioms of justice, from which "the law" in its entirety follows logically. Oliver Wendell Holmes argued pragmatically that "law and its institutions evolved from views of public policy, social context, history, and experience" (Minda, 1995:18) and that its application always relies on judgments about its role in society.

In educational and psychological testing, modernism corresponds to the pursuit of models and methods that characterize people through common variables, as evidenced by common observations—under the conceit that there are objectively correct ways of doing so. The source of these models and variables has been debated over the years along logical versus pragmatic

lines analogous to the Langdell versus Holmes stances. Witness on the one hand, factor analytic research programs that seek to "discover" the nature of intelligence and personality (e.g., Spearman, 1927; Thurstone, 1947) and, on the other, painstaking consen- sus-building procedures for assembling item pools to "measure achievement" in subject domains (e.g., Lindquist, 1951). These distinct branches within modern test theory correspond to the trait and behaviorist psychological perspectives. Under both perspectives, care is taken (1) to define, from the assessor's point of view, contexts in which to observe students; (2) to specify, from the assessor's perspective, the ways in which students' behavior will be summarized; and (3) to delineate the operations through which the assessor can draw inferences, within the assessor's frame of reference.

Postmodernism

> In jurisprudence, postmodernism signals the movement away from "Rule of Law" thinking based on the belief in one true "Rule of Law," one fixed "pat- tern," set of "patterns," or generalized theory of jurisprudence. . . . As developed in linguistics, literary theory, art, and architecture, postmodernism is also a style that signals the end of an era, the passing of the modern age . . . des- cribing what happens when one rejects the epistemological foundations of modernity. (Minda, 1995:224)

> Wittgenstein's view of language is that all of our language has meaning only within the language games and "forms of life" in which they are embedded. One must understand the use, the context, the activity, the purpose, the game which is being played. . . . (Minda, 1995:239)

The notion of discourse plays a central role in postmodernism. Language generates our "universe of discourse": the kinds of things we can talk about and the particular things we can say; what we construe as problems, how we attempt to solve them, and how we evaluate our success. But what is the source of words and concepts? Postmoderns claim that the commonsense idea that meanings of words reside "in" language is fundamentally misguided. For them, language constructs, rather than reflects, the meaning of things and events in the world (Minda, 1995:239).

Relativity and the quantum revolution shattered the belief that Newtonian and Euclidean models were the correct ultimate description of the universe. Ironically, improved instrumentation devised to finalize the modern research program revealed that its fundamental models were not in fact the universe's. Mathematical descriptions of observations departed increasingly from such intuitive notions as simultaneity and definitive locations of persistent entities. Just as ironically, while we obtain better accuracy in modeling phenomena and more power to solve applied problems than the "modern" physicists of the nineteenth century dreamed, we feel farther away from ultimate understanding. The universe is not only stranger than we imagine, mused the mathematician J.B.S. Haldane, it is stranger than we *can* imagine!

Just as relativity and quantum mechanics gave rise to postmodern physics, Minda noted several diverse movements that provoked a postmodern era in law in the 1980s: law and economics, critical legal studies, feminist legal theory, law and literature, and critical race theory. Cognitive psychology was the analogous shock to educational assessment—in particular, recognition of the crucial roles of students' perspectives in learning and of the social settings in which learning takes place. Snow and Lohman (1989:317) put it this way:

> Summary test scores, and factors based on them, have often been thought of as "signs" indicating the presence of underlying, latent traits. . . . An alternative interpretation of test scores as samples of cognitive processes and contents, and of correlations as indicating the similarity or overlap of this sampling, is equally justifiable and could be theoretically more useful. . . . Whatever their practical value as summaries, for selection, classification, certification, or program evaluation, the cognitive psychological view is that such interpretations no longer suffice as scientific explanations of aptitude and achievement constructs.

Neopragmatic Postmodernism

> Some postmodernists have adopted a *neopragmatic* outlook as an antidote to the postmodern condition. These postmodern critics are skeptical of the truth claims of modern theory, but they have not given up on theory. On the contrary, they believe that theory can have utility when used as a tool for the empirical investigation of problems. . . . Its

> practitioners accept the postmodern view that truth and knowledge are culturally and linguistically conditioned. On the other hand, neopragmatist practice is unlike . . . what some theorists call *poststructuralist* criticism because it is less concerned with exposing the contradictions of modern conceptual and normative thought than revealing instrumental, empirical, and epidemiological solutions for the problem at hand. (Minda, 1995:229-230)

Minda distinguishes "neopragmatic" postmodernists from "ironic" post- modernists, the latter of which "embrace the predicaments and paradoxes of the current intellectual condition in order to better understand the world of legal, social, and philosophical thought, and they attempt to bring out the irony of the experience of living in a postmodern world" (1995:4-5). In legal theory "the ironists attempt to facilitate the crisis and fragmentation of modern theory by employing postmodern criticism to 'displace, decenter, and weaken' central concepts of modern legal Western thought" (1995:230). In the fine arts, ironic postmodernism is rather de rigueur. Physics and, by extension, engineering demand a neopragmatic stance. Models and variables may indeed be our creations rather than nature's, and they are ever subject to alternatives and revisions—but we must in some way accommodate the constraints nature imposes upon us as we struggle with the challenges we confront. And if there is a job to do, languages, models, and conceptual frameworks are what we have to work with.

Like law, educational assessment lies somewhere between literature and physics. Cognitive research reveals recurrent patterns in the ways people learn and solve problems, yet what is important to learn and the conditions under which it will be learned are largely socially determined. "Neopragmatic postmodern test theory" explores the potential of using methodological and inferential tools that originated in a modern perspective to support learning in ways conceived in a postmodern perspective.

MODERN TEST THEORY

Technical Considerations

"Legal modernism also . . . is based on an understanding of language that assumes that words and conceptual ideas are capable of objectively capturing the meaning of events the law

seeks to describe and control" (Minda, 1995:6).

Most familiar practices of educational assessment can be traced to the first third of the twentieth century. Their forms were shaped by constraints on gathering and handling data in that era and by purposes conceived under then-current beliefs about learning and schooling. A paradigm of mental measurement analogous to classical (read "modern") physical measurement developed, and the tools of test theory evolved to guide applied work within this setting—designing tests, characterizing their evidential value, and evaluating how well they achieved their intended purposes. The targets of inference are aspects of students' learning, characterized as numbers on a continuum, upon which evaluations and decisions would be based if they were known with certainty.

In his 1961 article "Measurement of learning and mental abilities," Harold Gulliksen (1961:9) characterized the central problem of test theory as "the relation between the *ability* of the individual and his *observed score*." Referring explicitly to Plato's cave, he said "the problem is how to make the most effective use of these shadows (the observed test scores) in order to determine the nature of reality (ability) which we can only know through these shadows." The purposes of test theory, in this view, are to guide the construction of assessment elements and events (i.e., domains of test items and test conditions) and to structure inference from students' behavior in the resulting situations. The modernist underpinnings of the enterprise are reflected in a quotation from Gulliksen's (1961:101- 102) review of test theory on the occasion of the twenty-fifth anniversary of the Psychometric Society, concerning the search for the "right" item-response-theory models:

> An attempt to develop a consistent theory tying test scores to the abilities measured is typified by Lord's (1952) recent work . . . in which he formulated at least five different theories of the relationship between test scores and abilities, and showed how it was possible to test certain ones of these. It is to be hoped that during the next 10 or 20 years a number of these tests will be carried out so that we will have not five different theories of the relationship between ability and test score and various possible trace lines, but we will be able to say that, for certain specified tests constructed in this way, here is the relationship between the score and the ability measured, and this is the appropriate trace line to use.

Social Considerations

"Neopragmatists believe that theory merely establishes the rules for playing a particular language game" (Minda, 1995:236).

Although the physical measurement analogue connotes a certain objectivity and detachment, assessment based on the modernist approach nevertheless shapes, and is shaped by, social considerations. It structures conversations about learning in several ways:

- *Communication of expectations.* In and of themselves, domains of tasks and modes of testing convey, to students, teachers, and the public at large, what is important for students to learn and to accomplish.
- *Communication of results.* Once a domain of tasks and conditions of observation have been specified, a score and an accompanying measure of precision give a parsimonious summary of a student's behavior in the prescribed con- texts that is easily transmitted across time and place.
- *Credibility of results.* Test scores earn credibility beyond the immediate circumstances of the assessment if the data have been verifiably gathered under prescribed conditions.

That traditional assessment procedures serve these purposes is quite independent of the fact that they evolved under the mental measurement paradigm. Any procedures that might rise in their stead to assess and communicate students' learning would, in some way, need to address the same functions.

PROGENITORS OF CHANGE

The transition from the old to the 'new' jurisprudence began with the breakdown of the core beliefs and theories that served to define modern jurisprudence. The breakdown is partly a manifestation of the proliferation of new jurisprudential discourses and new movements in legal thought. (Minda, 1995:243)

I claimed earlier that developments in the psychology of learning and cognition brought about a postmodern era in assessment, and I shall say more about that later. These developments do indeed lay the groundwork for new developments in assessment, but I do not believe they were sufficient in and of them- selves to change the field. Had modern testing seen satisfactory progress in its research agenda, there would have been less impetus for change. But in assess-ment, as in physics, improved methodology and inferential methods led away from, rather than toward, the anticipated solutions.

Developments in Methodology

"There is a rising sentiment in the legal academy that modern legal theory has failed to sustain the modernists' hopes for social progress" (Minda, 1995:248).

Twenty-five years after Gulliksen's article, Charles Lewis observed that "much of the recent progress in test theory has been made by treating the study of the relationship between responses to a set of test items and a hypothesized trait (or traits) of an individual as a problem of statistical inference" (Lewis, 1986). New modeling and inferential techniques included item response theory, generalizability theory (Cronbach et al., 1972), structural equations modeling (e.g., Jöreskog and Sörbom, 1979), and the application of more powerful estimation methods from the statistical literature (e.g., Bock and Aitkin, 1981). They provided solutions to previously intractable problems such as tailoring tests to individual examinees and sorting out relationships in patterns of achievement test scores in hierarchical schooling systems.

These developments make for more efficient gathering of evidence and more powerful forms of argumentation for addressing questions that could be framed within the universe of discourse of modern test theory. But by requiring analysts to more clearly explicate their targets of inference and how observations provided evidence about them, these advances in modern test theory began to reveal important problems that lie beyond the paradigm's reach. The following two examples illustrate the point:

- How can we measure change, or can we? Through the use of standard test theory, evidence can be characterized and brought to bear on inferences about students'

overall proficiency in behavioral domains, for determining students' levels of proficiency, comparing them with others or with a standard, or gauging changes from one point in time to another. Cronbach and Furby (1970:76) cautioned that characterizations about the nature of this proficiency or how it develops fall largely outside the paradigm's universe of discourse:

> Even when [test scores] X and Y are determined by the same operation [e.g., a true-score or item-response-theory model for a specified domain of tasks], they often do not represent the same psychological processes. At different stages of practice or development different processes contribute to the performance of a task. Nor is this merely a matter of increased complexity; some processes drop out, some remain but contribute nothing to individual differences within an age group, some are replaced by qualitatively different processes.

- *Differential item functioning (DIF).* Classical test theory took test scores at face value, treating all response patterns with the same total score as identical. Item response theory explicated the conditions that would have to hold among patterns of item responses for total scores to capture all nonrandom variations among students. Essentially, the same expectation of success on each given task would have to hold for all students at a given true-score level, regardless of item content or students' background characteristics. Differential-item-functioning techniques devised to check these conditions often found that they failed in achievement tests—most importantly, in ways that related to curriculum (e.g., Miller and Linn, 1988) and solution strategies (e.g., Birenbaum and Tatsuoka, 1983; French, 1965). Because what is hard and what is easy is not universal— they depend, not surprisingly, on what and how students have been studying— summary scores inevitably fail to characterize some aspects of students' knowledge and progress.

Developments in Psychology

Cronbach and Furby's comments on measuring change presaged a growing awareness that

domain-referenced assessment methodologies, including item response theory, were simply not rich enough to support discourse about the nature and progress of students' learning. In assessment, as in physics, however, merely recognizing inadequacies in a paradigm is not sufficient for change. Newton and Huygens debated the contradictory wave- and particle-like properties of light as early as the seventeenth century. Paradigms are not displaced by data, the saying goes; paradigms are displaced only by other paradigms. Conceptions of learning that ground a broader universe of discourse for assessment are emerging from cognitive and educational psychology. The following paragraphs review some key insights into the ways people acquire and use knowledge and skills. Each, it will be noted, accents the uniquely personal and socially conditioned nature of learning.

- *Mental models/schema theory.* A "mental model" or "schema" is a pattern of recurring relationships—anything from what happens at birthday parties to how to figure out unit prices to how to carry out conversations—with variables that correspond to particular ways the pattern can occur. Some schemas are informal and intuitive; others we learn in part formally and explicitly. David Rumelhart (1980:55) claims that schemas

> play a central role in all our reasoning processes. . . . Once we can "understand" the situation by encoding it in terms of a relatively rich set of schemata, the conceptual constraints of the schemata can be brought into play and the problem readily solved.

No cognition is purely passive or data driven; we always construct meaning in terms of knowledge structures. Learning sometimes means adding bits to existing structures; sometimes it involves generalizing or connecting schemas; other times it involves abandoning important parts of schemas and replacing them by qualitatively different structures.

- *How expertise develops.* While experts in various fields of learning generally

command more facts and concepts than novices, the real distinction lies in their ways of viewing phenomena and representing and approaching problems (e.g., Chi et al., 1981). Experts learn to work from what Greeno (1989) calls the "generative principles of the domain," and they automatize recurring procedures (they "compile knowledge") so that they can devote their attention to novel aspects of problems. Increasing "metacognitive skills" also mark developing expertise: self-awareness in using models and skill and flexibility in how to construct them, modify them, and adapt them to problems.

- *Situated learning.* Assessment has focused on aspects of learning that are characterized insofar as possible as properties of individual students. Yet the nature of the knowledge we construct is conditioned and constrained by technologies, information resources, and social situations as we learn about tools, physical and conceptual, and how and when to use them. For example, reading comprehension depends on one's competence in recognizing words and parsing syntactic structures, but it also depends as much on am understanding of the context and substance of what the message is about. Students who have similar competences with structural aspects of language can take vastly different meanings away from the same text, depending on their experience with the phenomena in question. These findings, along with those discussed above, argue that learning is more richly characterized in terms of the student's breadth and configurations of connections across social and substantive contexts than by success in a given domain of tasks—even though such success occurs only by virtue of those connections.

These cross-cutting generalizations should not obscure the fact that cognitive psychology is a fractured, often fractious, field. Competing claims of rival researchers differ from one another as much as all differ from the trait and behaviorist perspectives. This is largely because different researchers are exploring different ranges of behavior, acquired and used under different circumstances. Birnbaum (1991:65) suggests:

Problem-solving depends on the manipulation of relatively fragmented and mutually

> inconsistent *microtheories*—each perhaps internally consistent, and each constituting a valid way of looking at a problem: "This will allow us to say, for example, that some [set of beliefs] is more appropriate than some [other set of beliefs] when confronted with problems of diagnosing bacterial infections. Scientists are used to having different—even contradictory—theories to explain reality. . . . Each is useful in certain circumstances." (Nilsson, 1991:45)

In assessment, as in law, the neopragmatic postmodernist welcomes all these lines of research as potentially useful tools for solving different practical problems; that is to say:

> For postmodern legal scholars, choosing the "best" answer for legal problems requires "tactical" judgments and questions regarding the values of the decision maker much more than a quest for a so-called "best" argument. One consequence of this has been the realization that there exists a multiplicity of answers for law's many problems. (Minda, 1995:252)

Rapprochement

Good teachers have always relied on a wider array of means to learn about how the students in their classes are doing and to help plan further learning. Alongside the tests and quizzes they design and score under the mental measurement paradigm, they also use evidence from projects, work in class, conversations with and among students, and the like—all combined with additional information about the students, the schooling context, and what the students are working on. Teachers call these "informal" assessments, in contrast with the "formal" assessments typified by large-scale standardized tests.

The stark contrast between formal and informal assessment arises because to understand students' learning and further guide it, teachers need information intimately connected with what their students are working on, and they interpret this evidence in light of everything else they know about their students and their instruction. The power of informal assessment resides in these connections. Good teachers implicitly exploit the principles of cognitive psychology,

broadening the universe of discourse to encompass local information and address the local problem at hand. Yet precisely because informal assessments are thus individuated, neither their rationale nor their results are easily communicated beyond the classroom. Standardized tests do communicate efficiently across time and place—but by so constraining the universe of discourse that the messages often have little direct utility in the classroom.

The challenge now facing neopragmatic postmodern test theory is to devise assessments that, in various ways, incorporate and balance the strengths of formal and informal assessments by capitalizing on an array of methodological, techno- logical, and conceptual developments.

POSTMODERN TEST THEORY

"Postmodern legal critics employ local, small-scale problem-solving strategies to raise new questions about the relation of law, politics and culture. They offer a new interpretive aesthetic for reconceptualizing the practice of legal interpretation" (Minda, 1995:3).

Cognitive psychology challenges the adequacy of the "one-size-fits-all" presumption of standard assessment, which defines the target of inference in terms of an assessor-specified domain of tasks, to be administered, scored, and interpreted in the same way for all students. The door has been opened to alternative ways to characterize students' proficiency and acquire evidence about it—ways that may involve observing students in different situations, interpreting their actions in light of additional information about them, or triangulating across context and situation, as may be required for one's purpose (Moss, 1996).

Moss (1994) and Delandshere and Petrosky (1994) offer postmodern insights into assessment from a less structural perspective than mine, criticizing test theory as it is conceived from a modernist point of view. I am interested in the utility of model-based inference in assessment, as *re*conceived from a postmodernist point of view. I submit that concepts from psychology and inferential tools from model-based reasoning can support assessment practice as more broadly conceived—just as Newton's laws still guide bridge design, quantum mechanics and relativity theory notwithstanding. The essential elements of the approach are (1) understanding the important concepts and relationships in the learning area in order to know the

aspects of students we need to talk about in the universe of discourse our assessment will generate and (2) determining what one needs to observe and how it depends on students' understandings, so as to structure assessment settings and tasks that will provide evidence about the above- mentioned aspects of students' understandings. Here is an example from a project I have been working on recently, concerning advanced placement (AP) studio art portfolios.

Viewed only as measurement, the AP studio art portfolio program would be a disaster. Students spend hundreds of hours creating the portfolios they submit for scoring at the end of the year, and raters who are art educators and teachers spend hundreds of hours evaluating the work—all to produce reliability coefficients about the same as those of 90-minute multiple-choice tests. The situation brightens when the program is viewed as a framework for evidence about skills and knowledge, around which teachers build art courses with wide latitude for topics, media, and projects. A common understanding of what is valued and how it is evaluated in the central scoring emerges through teacher workshops, talked- through examples with actual portfolios, and continual discussions about how to cast and apply rating rubrics to diverse submissions. Meaning emerges through countless conversations across hundreds of classrooms, each individual but with some common concepts and shared examples of their use—each enrichened and individuated locally in a way that grounds instruction and local evaluations but with a common core that grounds more abbreviated program-wide evaluations. This is, at heart, a social phenomenon, not a measurement phenomenon. Carol Myford and I have found an item-response-theory measurement model for ratings valuable, nevertheless, to illuminate how raters use evaluative criteria and to characterize uncertainty about students' scores (Myford and Mislevy, 1995). We do not use the model to "gauge the accuracy of a measuring instrument." We use it to survey patterns of similarity and variation, of agreement and disagreement, among tens of thousands of virtual dialogues among students, raters, and teachers, through their portfolios—to the end of discovering sources of misunderstanding and cross talk that can frustrate the conversations.

Model-based reasoning is useful not so much for characterizing the unique essence of a phenomenon but as a tool of discourse—for organizing our thinking, for marshaling and interpreting evidence, and for communicating our inferences and their grounding to others. The discipline that model-based reasoning demands even benefits us when we don't believe the models are true: it is easier to notice phenomena that don't accord with the patterns we expect to

see and, therefore, to revise our thinking. A skeptical attitude about models in assessment makes our uses of them more flexible, more powerful, and, ultimately, more effective at meeting and fulfilling the aims of education than they would be if we believed that they accurately captured the totality of the phenomenon.

From a modernist perspective, Lord Kelvin declared at the turn of the century that "when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind." Measurement, in his eyes, was a one-off representation of truth. From a postmodern perspective, even if you can measure, your knowledge is still meager in a fundamental sense but at least you have a practicable framework for discourse—for structuring action, for communicating your observations and your reasoning, for struggling with practical problems, for surprising yourself in ways that lead to further understanding. Lord Kelvin's quote is a modernist scientific version of Yogi Berra's "it ain't over 'til it's over." The postmodern response is Jesse Jackson's "and even then it ain't over."

SOME IMPLICATIONS OF A POSTMODERN PERSPECTIVE

> Neopragmatists thus attempt to explain how one can do theoretical work with- out rejecting all pretenses of foundational knowledge. Neopragmatists argue that the theorists must take a situated stance in their scholarship and adopt an instrumental approach to theory. Whatever works in context becomes the standard for their theoretical investigation and judgment. . . . When applied to legal studies, neopragmatism forms the academic perspective of scholars who reject all foundational claims of legal theory but remain committed to the view that legal theory can be useful for solving legal problems. . . . Neopragmatists thus believe in and are committed to the Enlightenment idea of progress, even while they resist using the modernist's framework. (Minda, 1995:230)

In the remainder of the main body of the paper, I offer comments from a neopragmatic postmodern perspective on enhancing familiar kinds of assessment, even while moving our interpretational perspective beyond its modernist roots. As an example, I address the question of the degree to which "adult literacy," an essential element of workplace skill, can be defined and

gauged across literacy training programs.

Progress Within Modern Test Theory

Familiar forms of assessment were shaped by constraints on how data could be gathered, stored, transmitted, and analyzed. Logistical and economic pressures limited the large-scale use of essays and interviews that required human interpretation, thus favoring objective-response tasks over more constructive and sustained tasks. It was not possible to store or share ephemeral performances in order to develop common standards or to verify that ratings were fair. These constraints are being eased by technological developments—computers, video- taping and audio-taping, electronic communication, mass storage, and access to resources. Some new possibilities appear even within the traditional mental measurement paradigm. I will mention some briefly but then argue that technology alone will not break through the essential inferential barriers of the modernist test theory perspective:

- *New kinds of tasks and scoring*. Computers can present students with tasks that are interactive (e.g., simulated experiments), dynamic (e.g., medical treatment problems in which simulated patients' conditions change over time), constructive (e.g., a stimulated construction site onto which elements must be moved to meet client's needs), and less tightly structured (e.g., a word problem that is to be approached in many ways). Some scoring can be also done automatically, including for these examples.

- *Distributed testing and scoring.* Students' responses to computerized tasks can now be captured and electronically transmitted. Performances can be videotaped and audio- taped. Constructed paper-and-pencil responses and artwork can be scanned. Students can thus be assessed in remote places and at different times, and raters can evaluate their performances in remote places and at different times. Students in school consortiums can share work on a common project, interacting with, and receiving feedback from, teachers and students across the nation.

- "*Replayability*" (Frederiksen and Sheingold, 1994). Beside easing constraints on time and

> location, capturing performances helps address the rater- agreement problems that troubled Horace Mann more than a century ago. Performances can now be seen, discussed, and evaluated by as many people, in as many times and places, as desired. Now that we are no longer limited to the evaluations of raters present at the original performance, we can have broader and more interconnected scoring of individual students and use exemplars to establish shared expectations and standards of evaluation, over time and across distance, among raters, teachers, and students.

Despite technology and efficient statistical models, the objective of characterizing students' proficiency must remain poorly met if one is constrained to one-size-fits-all data and to ignorance of contextual and educational background factors. The more examinees differ as to relevant contextual and experiential factors, the more likely it is that each task in a complex and context-rich domain will consume considerable time and costs without providing much information about how students would fare on other tasks—the Shavelson et al. (1992) "low generalizability" problem (also see Linn, 1993). Each individual task may provide copious information for some inferences—but not for inferences about the usual target, domain-true score. The same complex task can be invaluable in an assessment linked with instruction and grounded in context yet worthless in a broadly cast survey because it is trivial, unapproachable, or incomprehensible to most students.

Beyond Modern Test Theory

> Postmodernism thus challenges legal thinkers to reconsider their most basic understanding of the nature of law and politics—their belief in an objective and autonomous law. Postmoderns argue that decision making according to rule is not possible, because rules are dependent upon language, and language is socially and culturally constructed and hence incapable of directing decision makers to make consistent and objective choices. Objectivity is possible only if agreement or consensus about different interpretive practices can be reached. (Minda, 1995:245)

Standards of content and performance are a topic of intense current interest. I have argued that limited sets of common assessment tasks, scored and interpreted in common ways that

ignore context, cannot by their nature tell us all we would like to know about students' learning. They may tell us something worth knowing, especially if the inferences and actions based on them do take context into account (Messick, 1989). As noted in the preceding section, technology is broadening the span and efficiency of such assessment. And with such assessment it is possible to gauge the levels of performance of individual students and groups of students. The real challenge, it seems, is to extend the notion of standards beyond the confines of the modernist perspective: Is it possible to retain the relevance and connectedness traditionally associated with informal assessment yet simultaneously serve the communicative and credibility-based functions traditionally associated with formal assessment?

The AP studio art experience suggests that the answer is yes. Learning there is individuated, but a shared conception of the nature of intended learning, developed through examples and feedback, makes it possible to interpret work in a common framework. Such a structure appears necessary if assessments with constructive and individuated data, such as portfolios and exhibitions (e.g., Wiggins, 1989), are to span time and distance.

Common meaning is necessary for credibility, but it is not sufficient. Why should anyone trust an interpreted evaluation of a performance from a distant time and place? Standardized test results gain a measure of credibility from their prescribed procedures; these are established "rules of the game," which, if followed, circumscribe the interpretation of the results. Even though the results don't tell about everything that is important, parents and boards of education can ask questions and verify procedures in order to spot invalidating practices. But the more individuated an assessment is, the more difficult it becomes to establish credibility.

For example, in some ways teachers are in the best position to evaluate students' work, by virtue of their knowledge about context and situation. Their contextualized evaluations are unquestionably basic for guiding classroom learning. Can their evaluations be used for high-stakes purposes beyond the classroom, in light of their vested interest in their students' success and the typically wide variation in their interpretations of performance? As noted above, a common framework for interpretation is required first. The validity of mappings of performances into that framework can be addressed by mechanisms such as audits, cross evaluation across schools, and triangulation of types of evidence (Resnick, in press). Technology can play an enabling role, through replayability, mass storage, and electronic communication. Statistical modeling can play a quality assurance role, through the analysis of ratings of multiply-scored

work, as discussed above in connection with AP studio art.

An Example: Adult Literacy Assessment

As defined by the National Literacy Act of 1991, literacy involves "an individual's ability to read, write, and speak in English, compute, and solve problems at levels of proficiency necessary to function on the job and in society, to achieve one's goals, and to develop one's knowledge and potential." The act requires state education agencies to "gather and analyze data—including standardized test data—on the effectiveness of State-administered adult education programs, services, and activities, to determine the extent to which the State's adult education programs are achieving the goals in the state plan" [to enhance levels of adult literacy and improve the quality of adult education services] (*Federal Register,* 1992). These federal evaluation requirements have prompted interest in identifying standardized tests and methodologies that are appropriate for assessing the effectiveness of adult education programs and for determining the feasibility of linking such tests in order to provide national trend data on program effectiveness (Pelavin Associates, 1994).

But the diversity of both the objectives and the participants served by adult education programs reflects a broad and multidimensional definition of literacy. Accordingly, adult education programs vary considerably with respect to the nature and level of skills they emphasize and with respect to the kinds of students with whom they work. Some strongly resemble and largely replace the academic reading experience that high schools supply, in order to help dropouts obtain General Education Development certificates. Others help immigrants and others who are literate in languages other than English to speak, read, and write in English. Still others work with adults who are literate, if not skilled, from the perspective of traditional schooling, in order to develop more specific skills for use in the workplace. Moreover, these diverse programs use tests for a broad variety of diagnostic, instructional, and evaluative purposes. Both the nature of the instruction and the purpose of testing determine the kinds of tests that will be appropriate.

Is it possible to link results from these varied tests, across the diverse pro- grams, to secure a common metric for evaluating program effects and tracking trends over time? Writing on the prospect of calibrating disparate tests to common national standards, Andrew Porter (1991:35)

wrote,

If this practice of separate assessments continues, can the results be somehow equated so that results on one can also be stated in terms of results on the other? There are those who place great faith in the ability of statisticians to equate tests, but that faith is largely unjustified. Equating can be done only when tests measure the same thing.

Professor Porter's skepticism is justified. We are perhaps too familiar with correspondence tables that give exchangeable scores for alternate forms of standardized tests. But they work only because the alternate forms were constructed to meet the same tight specifications; equating studies and statistical formulas merely put into usable form the evidentiary relationships that were built into the tests (see Linn, 1993, and Mislevy, 1993, for definitions, concepts, and approaches that have been developed to link educational tests for various purposes).

Statistical procedures neither create nor determine relationships among test scores. Rather, the way that tests are constructed and administered and the ways that the skills they tap relate to the people to whom they are administered determine the nature of the potential relationships that exist in evidence that scores from the various tests convey. Much progress has been made recently with statistical machinery for this purpose, with power beyond the expectations of educational measurement researchers a generation or two ago. However, we now recognize the objective of building once-and-for-all correspondence tables as a chimera—it is not simply because we lack the tools to answer the question but because the question itself is vacuous. Statistical procedures, properly employed, can be used to explicate the relationships that do exist in various times and places and harness the information they do convey for various purposes. Perhaps more importantly, they help us understand what information different tests do not, indeed cannot, convey for those purposes.

Thus, the first two conclusions listed below, about what can be expected from applying statistical linking procedures to adult literacy tests, are negative. They repudiate a naive modernist goal of rectifying various indicators of a common true variable, when those indicators have evolved to serve different purposes in different contexts, gathering qualitatively different kinds of information.

- *No single score can give a full picture of the range of skills that are important to all the different students in different adult literacy programs.*

- *No statistical machinery can translate the results of any two arbitrarily selected adult literacy tests so that they provide interchangeable information about all relevant questions about student competencies and program effectiveness.*

What *is* possible? Three less ambitious, but more realistic, affirmative contingencies, each employing modernist statistical techniques from a neopragmatic postmodernist perspective. All require the prerequisite realization that no test scores can capture the full range of evidence about students' developing proficiencies within their courses, nor can they convey all that is needed to determine how well students are progressing toward their own objectives. This understood, here are some options for dealing with such information as there is in literacy test scores, when different literacy programs must use different tests in accordance with their differing goals and instruction:

- *Comparing directly the levels of performance across literacy programs in terms of common indicators of performance on a market basket of consensually defined tasks in standard conditions.* Some aspects of competence, and assessment contexts for gathering evidence about them, will be considered useful by a wide range of programs, and components of an assessment system can solicit information about them in much the same way for all. However, these "universal" assessments—and in particular pre-post comparisons with such assessments—provide seriously incomplete information to evaluate the effectiveness of programs, to the extent that their focus does not match the programs' objectives (to say nothing of the *students'* objectives!).

- *Estimating levels of performance of groups or individuals within clusters of literacy programs with similar objectives—possibly in quite different ways in different clusters—at the levels of accuracy demanded by purposes within clusters, with shared assessments focused on those objectives.* These components of programs' assessments might gather evidence for different purposes, types of students, or levels of proficiency, to complement information gathered by "universal" components.

- *Making projections about how students from one program might have performed on the assessment of another.* When students can be administered portions of different clusters' assessments under conditions similar to those in which they are used operationally, the joint distribution of results on those assessments can be estimated. These studies are restricted as to time, place, program, and population, however. The more the assessments differ as to their form, con- tent, and context, the more uncertainty is associated with the projections, the more they can be expected to vary with students' background and educational characteristics, the more they can shift over time, and the more comparisons of program effects become untrustworthy.

CONCLUSION

It is a critical time for jurisprudential studies in America. It is a time for self- reflection and reevaluation of methodological and theoretical legacies in the law. At stake is not only the status of modern jurisprudence, but also the validity of the Rule of Law itself. In the current era of academic diversity and disagreement, the time has come to seriously consider the transformative changes now unfolding in American legal thought. The challenge for the next century will certainly involve new ways of understanding how the legal system can preserve the authority of the Rule of Law while responding to the different perspectives and interests of multicultural communities. It is without a doubt an anxious and exciting time for jurisprudence. . . .

> What was once understood as the mainstream of modern view has broken into a diverse body of jurisprudential theories and perspectives. . . . No matter how troubling it may be, the landscape of the postmodern now surrounds us. It simultaneously delimits us and opens our horizons. It's our problem and our hope. (Minda, 1995:256-257)

Ironist critics of educational assessment reject the modernist notion that the "truth" about a student's understanding or a program's effect lies but a simple step away from our ken, to be spanned by observations with standard, context-free measuring instruments and unambiguous statistical analysis of the results. But to further reject any use of these models and information-gathering tools just because they arose under the discarded epistemology is to forgo decades of

experience about some ways to structure and communicate observations about students' learning. Educators fear that wholesale abandonment of familiar assessment methodology strips away tools that help them address these facets of their task. Believing these ways of structuring discourse hold *no* value is as wrong as believing that they *alone* hold value. I hear parents and teachers say that we "should not throw the baby out with the bath water." But how to tell which is which?

My answer (a neopragmatic postmodernist answer, as it turns out) is this: Models, principles, and conceptual frameworks are practicable tools—not for discovering a singular truth but for structuring our discourse about students, so that we may better support their learning, and for learning about expected and unexpected outcomes of our efforts, so that we may continually improve them. Understandings of students' learning and programs' effects are enriched by multiple perspectives and diverse sources of evidence, some new or previously neglected but others with familiar (albeit reconceived) forms. Postmodern architects play with ironies in design, advancing alternative sensibilities and forgotten voices—but they had better design buildings that are livable and safe. Fundamental constraints and fundamental responsibilities persist. And as long as we in education purport to help other people's children learn, at other people's expense, we bear the duty of gaining and using as broad an understanding as we can to guide our actions and of conveying our reasoning and results as clearly as we can to those to whom we are responsible.

ACKNOWLEDGMENTS

I am indebted to Bob Linn and Barbara Storms for their thoughtful comments on an earlier version.

References

Birenbaum, M., and K.K. Tatsuoka. (1983). The effect of a scoring system based on the algorithm underlying the students' response patterns on the dimensionality of achievement test data of the problem solving type. *Journal of Educational Measurement* 20:17-26.

Birnbaum, L. (1991). Rigor mortis: A response to Nilsson's "Logic and artificial intelligence." *Artificial Intelligence* 47:57-77.

Bock, R.D., and M. Aitkin (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. *Psychometrika* 46:443-459. Chi, M.T.H., P.

Feltovich, and R. Glaser (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science* 5:121-152.

Cronbach, L.J., and L. Furby (1970). How should we measure "change"—Or should we? *Psychological Bulletin* 74:68-80.

Cronbach, L.J., G.C. Gleser, H. Nanda, and N. Rajaratnam (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: Wiley.

Delandshere, G., and A.R. Petrosky (1994). Capturing teachers' knowledge: Performance assessment (a) and poststructural epistemology, (b) from a post-structuralist perspective, (c) and post-structuralism, (d) one of the above. *Educational Researcher* 23(5):11-18.

Frederiksen, J.R., and K. Sheingold (1994). *Linking Assessment with Reform: Technologies that Support Conversations About Student Work* Princeton, NJ: Center for Performance Assessment, Educational Testing Service.

French, J.W. (1965). The relationship of problem-solving styles to the factor composition of tests. *Educational and Psychological Measurement* 25:9-28.

Greeno, J.G. (1989). A perspective on thinking. *American Psychologist* 44:134-141.

Gulliksen, H. (1961). Measurement of learning and mental abilities. *Psychometrika* 26:93-107.

Jöreskog, K.G., and D. Sörbom (1979). *Advances in Factor Analysis and Structural Equation Models.* Cambridge, MA: Abt Books.

Lewis, C. (1986). Test theory and *Psychometrika*: The past twenty-five years. *Psychometrika* 51:11-22.

Lindquist, E.F. (1951). Preliminary considerations in objective test construction. Pp. 119-185 in *Educational Measurement*, E.F. Lindquist, ed. Washington, DC: American Council on Education.

Linn, R.L. (1993) Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis* 15:1-16.

Lord, F.M. (1952). A theory of test scores. *Psychometrika Monograph* 17(4, Part 2):1-80.

Messick, S. (1989). Validity. Pp. 13-103 in *Educational Measurement*, 3rd ed, R.L. Linn, ed. New York: American Council on Education/Macmillan.

Miller, M.D., and R.L. Linn (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement* 25:205-219.

Minda, G. (1995). *Postmodern Legal Movements.* New York: New York University Press.

Mislevy, R.J. (1993). *Linking Educational Assessments: Concepts, Issues, Methods, and Prospects.* Princeton, NJ: Educational Testing Service.

Myford, C.M., and R.J. Mislevy (1995). *Monitoring and Improving a Portfolio Assessment System.* Princeton, NJ: Center for Performance Assessment, Educational Testing Service.

Moss, P. (1994). Can there be validity without reliability? *Educational Researcher* 23(2):5-12. (1996) Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher* 25(1):20-28. Nilsson, N.J. (1991) Logic and artificial intelligence. *Artificial Intelligence* 47:31-56. Pelavin Associates (1994) *Comparing Adult Education Tests: A Meeting of Experts.* Washington, DC: Pelavin Associates.

Porter, A.C. (1991) Assessing national goals: Some measurement dilemmas. Pp. 21-42 in *The Assessment of National Goals. Proceedings of the 1990 ETS Invitational Conference,* T. Wardell, ed. Princeton, NJ: Educational Testing Service.

Resnick, L. (1994). Performance puzzles. *American Journal of Education* 102(4):511-526.

Rumelhart, D.A. (1980) Schemata: The building blocks of cognition. Pp. 33-58 in *Theoretical Issues in Reading Comprehension*, R. Spiro, B. Bruce, and W. Brewer, eds. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shavelson, R.J., G.P. Baxter, and J. Pine (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher* 21(4):22-27.

Snow, R.E., and D.F. Lohman (1989). Implications of cognitive psychology for educational measurement. Pp. 263-331 in *Educational Measurement*, 3rd ed, R.L. Linn, ed. New York: American Council on Education/Macmillan.

Spearman, C. (1927). *The Abilities of Man: Their Nature and Measurement.* New York: Macmillan.

Thurstone, L.L. (1947). *Multiple-Factor Analysis*. Chicago: University of Chicago Press.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan* 70:703-713.