ASSESSING WRITING, 1 (1), 109–128, (1994)

Validity in High Stakes Writing Assessment: Problems and Possibilities

PAMELA A. MOSS
University of Michigan

For many years now, there has been a productive tension between the disciplines of educational measurement and literacy education. Some of the recent developments in assessment—especially the move toward performance-based and portfolio assessments—received their first extensive trials in the context of writing assessment. Initially, educators' concerns about the quality of information provided by multiple-choice or "indirect" measures of writing ability and about the consequences of using such assessments for teaching and learning led to the call for more direct writing assessments involving actual samples of writing. Extensive research has been conducted on how to develop and score standardized writing tasks to provide reliable, valid, and fair estimates of students' writing abilities (e.g., Breland, Camp, Jones, Morris, & Rock, 1987; Huot, 1990; Ruth & Murphy, 1988). As of 1991, 36 states were using tests that included writing samples and nine others had them under development (Office of Technology Assessment [OTA], 1992). Participation in the design, scoring, and use of these assessments provided opportunity for professional dialogue among teachers (e.g., OTA, 1992). Renewed concerns about the quality of information and consequences from these standardized writing assessments—typically first drafts completed in brief amounts of time—taken together with new understandings about the cognitive and social aspects of learning, led to the call for more complex and authentic writing assessments. Such assessments should provide students with the opportunity to explore more of their own purposes, to rethink and revise their work over extended periods of time, drawing on existing resources and responses from readers, and to reflect on the process and quality of their writing (e.g., Camp, 1992a; Wolf, Bixby, Glenn, & Gardner, 1991). Pilot testing of extended performance and/or portfolio assessments is well underway in a few states (see

I am grateful to Roberta Camp and Caroline Taylor Clark for thoughtful comments on an earlier draft of this article.

Correspondence and requests for reprints should be sent to Pamela A. Moss, University of Michigan, 4220 School of Education, Ann Arbor, MI 48109-1259.

OTA, 1992 for brief descriptions and references). Teachers participating in the development, scoring, and use of these assessments are again benefiting from the collegial dialogues around student writing which have led to re-examination of the goals, activities, and standards for learning (e.g., Camp, 1992b; Koretz, Stecher, & Diebert, 1992). Measurement researchers are exploring ways to develop large scale extended performance and portfolio assessments that both encourage authentic work and that can be used with reliability, validity, and fairness to inform consequential decisions about individuals and programs (e.g., Koretz, McCaffrey, Klein, Bell, & Stecher, 1992; Linn, Baker, & Dunbar, 1991; Wiley & Haertel, in press). It is here, however, that the tension between the disciplines of educational measurement and literacy education is at risk of becoming less productive. Experience suggests that in order to achieve the standards of validity necessary for high stakes purposes—for informing consequential decisions about individuals and programs—assessments need to be standardized to some degree. Standardization refers to the extent to which tasks, working conditions, and scoring criteria are similar for all students. Emerging views of literacy, however, suggest the need for less standardized forms of assessment to support and document purposeful, collaborative work by students and teachers. This results in the tension between competing validity criteria that simultaneously advocate standardization and purposeful, collaborative activity. Proposed solutions often reflect compromises between competing criteria rather than the kind of fundamental rethinking that might push both fields forward. It is the set of problems and possibilities contained in this tension that I want to explore in this paper.

THE GENESIS OF THE TENSION

Increasingly, writing and other forms of literacy are coming to be viewed more as purposeful, meaningful activities than as abilities. John Willinksy (1990), in his discussion of the "new literacy," offers an analogy between literacy and bicycle riding.

> The point is not to develop the ability to ride, which leads to sessions of practicing and demonstrating the skill.... If bikes are worth riding then the learning should begin with the intent of taking you places.... What is important about riding are the places to which you ride and the pleasures gained along the way. In the process of this riding with a purpose, the skill naturally improves. (p. 8)

Similarly, Maxine Greene raises the concern that evaluation which foregrounds testable skills may well prevent students "from taking responsibility for their own questions, their own learning to learn" (p. 11). When

reading and writing are viewed as purposeful activities, it becomes crucial to provide students with some opportunity to negotiate their own purposes and practices beyond simply demonstrating their competence in school (Newmann, 1990). And, it becomes necessary to provide teachers with the opportunity to mediate the delicate balance between structure and freedom in facilitating and evaluating their students' learning (Darling-Hammond & Snyder, 1992).

Recent developments in the philosophy of validity lend theoretical support to this direction. Most validity theorists (e.g., Cronbach, 1988, 1989; Messick, 1989, 1992) have argued for expanding the concept of validity beyond its traditional focus on the soundness of inferences about students' capabilities (construct validity) to include explicit consideration of the intended and unintended consequences of using an assessment (consequential validity).[1] Increasing evidence about the impact of assessment on teaching and learning suggests the importance of assessing the full range of capabilities and interests we want to nurture in our students. Validity researchers in performance assessment have stressed the importance of *balancing* traditional validity concerns about reliability, comparability, and generalizability with additional criteria such as "authenticity" (Newmann, 1990), "directness" (Frederiksen & Collins, 1989), or "cognitive complexity" (Linn, Baker, & Dunbar, 1991), as long as "acceptable levels are achieved for particular purposes of assessment" (Linn, Baker, & Dunbar, 1991, p. 11).

Less standardized forms of assessment, such as performance assessments, however, have presented serious problems for reliability. Reliability refers to consistency, quantitatively defined, among measures which are intended as *interchangeable*—consistency among independent evaluations or readings of a performance, consistency among performances in response to independent tasks, and so on (AERA, APA, NCME, 1985; Feldt & Brennan, 1989). In current validity theory, reliability is necessary to support inferences from particular samples of work evaluated by particular readers to the broader capabilities the assessments are intended to tap. Empirical studies of reliability or generalizability with performance assessments are quite consistent in their conclusions that (a) reader reliability, defined as consistency of evaluation across readers on a given task, can reach acceptable levels when carefully trained readers evaluate responses to one task at a time, but that (b) adequate task or "score" reliability, defined as consistency in performances across tasks intended to address the same capabilities, is far

[1]Moss (1992) and Shepard (1993) provide overviews of recent developments in the philosophy of validity beyond what is reflected in the most recent *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1985), which is under revision.

more difficult to achieve (e.g., Breland et al., 1987; Dunbar, Koretz, & Hoover, 1991; Shavelson, Baxter, & Gao, 1993). In the case of portfolios, where the tasks may vary substantially from student to student and where multiple tasks may be evaluated simultaneously, inter-reader reliability may drop below acceptable levels for consequential decisions about individuals or programs (Koretz, McCaffrey, Klein, Bell, & Stecher, 1992; Nystrand, Cohen, & Dowling, 1993). Recommendations for enhancing reliability, without increasing the number of tasks or readers beyond cost-efficient levels, have typically involved increasing the degree of standardization in one or more aspects of assessment. This sets up a tension between two very real and appropriate concerns—between assessments that promote and document a purposeful and collaborative view of literate activity and assessments that reflect adequate levels of construct validity (and reliability) for high stakes decisions about individuals and programs.

A seemingly simple solution to this tension between standardization and purposeful activity is to suggest that less standardized assessments be used at the classroom level for monitoring student progress and that the more standardized forms of assessment be used for high stakes purposes such as individual placement, selection, and certification and/or program evaluation and accountability. This suggestion has been made both by psychometricians (e.g., Mehrens, 1992) and literacy educators (e.g., Tavalin, 1993) who have felt the constraints of meeting competing criteria. However, given the well-documented power that externally imposed assessments have on teaching and learning, this solution is problematic—it is at best optimistic to think that the assessments "that count" will not overpower the alternatives in the lives of teachers and students.

PAST EXPERIENCE WITH CONSEQUENCES OF HIGH STAKES TESTING[2]

Consider what we have learned from over a decade of experience with high stakes, primarily multiple choice assessment. A growing body of evidence indicates that when assessments are visible and have consequences for individuals or programs, they alter educational practice, sending an unequivocal message to teachers and students about what is important to teach and learn (Madaus, 1988; National Commission on Testing and Public Policy, 1990; Resnick & Resnick, 1992). In a review of literature on the impact of classroom evaluation on students, Crooks (1988) concluded assessment not only affects students' judgments about what is important to learn, but also their motivation, perceptions of competence, approaches to

[2]This section is adapted from Moss and Herter (1993).

personal study, and development of enduring learning strategies. Similar conclusions have been drawn about the impact of district and state mandated assessment on the judgment, perceptions, and instructional strategies of teachers. The salience of this influence seems to be directly related to the importance of the consequences of testing to students and teachers and to the administrative and supervisory practices of a school or district. In a paper prepared for the National Commission on Testing and Public Policy, Resnick and Resnick concluded that "when the stakes are high—when schools' ratings and budgets or teachers' salaries depend on test scores—efforts to improve performance on a particular assessment seem to drive out most other educational concerns" and "to progressively restrict curricular attention to the objectives that are tested and even the particular item forms that will appear on the test." (1992, p. 58).

For example, in case studies of two public schools, Mary Lee Smith (1991) observed that high stakes tests not only narrowed the curriculum in subjects tested but also led to the neglect of untested subjects. Moreover, she noted that these effects were not just seasonal, but had long term consequences for curriculum development. In one school, teachers pushed to eliminate the hands on science curriculum and the writing process curriculum because they did not coordinate with the district's testing program. Research by Peter Johnston and colleagues (Johnson, Weiss, & Afflerbach, 1990), suggests that there are more subtle effects on teachers' perceptions of their students. They interviewed over 50 English/Language Arts teachers from 5 different districts that ranged in degree of prominence placed on externally imposed tests. Among other questions, teachers were asked to describe their students' literacy development. Based upon their responses, Johnston and his colleagues concluded that when tests were emphasized by the district, teachers tended to describe their students' achievements in terms of tests, competitive attainments, and test-like language, whereas when literature was emphasized in the classroom, teachers tended to describe students in terms of the books they had chosen to read, their written reflections on those books, and their individualized progress through literature.

Evidence suggests that the narrowing of the curriculum associated with high stakes standardized assessment may be falling disproportionately on certain groups of students for whom concerns about equality of education have been most salient. Consider, for instance, what we know about the impact of assessment on low income, urban students. Neill and Medina (1989) found that standardized testing was more prevalent in large urban school systems and the National Assessment for Educational Progress reported that students attending schools in and around large cities, where a high proportion of residents are on welfare or not regularly employed, scored lower on achievement tests than students from other types of

American communities (U.S. Department of Education, 1988, 1989). Herman and colleagues (Herman & Golan, 1993; Dorr-Bremme & Herman, 1986), using teachers' and principals' self-reports, found that in low income communities, teachers felt a greater need to spend time preparing students for tests and principals felt that tests counted far more in decisions such as planning curriculum, making class assignments, allocating funds, and reporting to district officials and the community. To the extent that testing undergirds decisions about educational placement, studies on the effects of tracking reviewed by Oakes, Gamoran, and Page (1992) also support concerns about the differential impact of testing on low income, urban students. They report that tracks for low ability and non-college bound students have higher proportions of low-income students; that qualitative differences exist in the educational experiences provided students in different tracks, with lower track students progressing more slowly through the curriculum, having less experience with inquiry skills, problem solving, and autonomy in their work, and losing more educational time to classroom management; that the achievement gap between students in higher and lower tracks increases over years of schooling; and that track placement can have a long lasting impact on the life chances of students after high school. Taken together with our knowledge about the impact of high stakes testing on the curriculum, these observations raise substantial concerns about differential access to knowledge for low-income urban students and for other groups of concern.

Although high stakes testing programs frequently result in improved test scores, such improvement does not necessarily imply a rise in the quality of education or a better educated student population (Darling-Hammond & Snyder, 1992; Haertel, 1989; National Commission on Testing and Public Policy, 1990; Shepard, 1992). At best, test scores can reflect only a small subset of valued education goals. Historically (and appropriately) tests have been considered indicators of educational achievement, and their validity has rested on their relation with more direct measures of the capabilities they are intended to predict. When educators focus their attention on improving test scores, they not only narrow the curriculum; they undermine the validity of the tests as indicators of a broader range of achievements (Shepard, 1992).

Further, evidence suggests that test-driven reforms may undermine attempts at genuine educational reform by diverting attention from fundamental educational problems. Ellwein, Glass, and Smith (1988) conducted extended case studies of five competency testing programs at the state and district level. They concluded that competency tests and standards served more as symbolic and political gestures rather than as instrumental reforms—focusing attention on the tests themselves rather than on their impact, utility, or value. Similarly, Corbett and Wilson (1991), who studied

competency testing programs in two states, focusing on six districts per state, found that the pressure to do well on tests did not encourage fundamental consideration of the structures, processes, or purposes of education; rather it caused "knee-jerk" reactions designed to improve test scores quickly—actions which many of the educators involved considered counter-productive.

ALTERNATIVE MODELS FOR HIGH STAKES ASSESSMENT

Some policymakers have attributed the problems encountered with high stakes testing to the "low-level" outcomes typically assessed and the lack of coordination between assessments and curriculum. In this view, the solution to the problems encountered rests in designing assessments that are closely integrated with curriculum standards and that encompass a wider range of valued educational goals. As Linn (1993) notes, these assumptions are reflected in a number of recent reform proposals (see also, Baker, 1989). Such assessments, it is hoped, will not only permit valid inferences about the quality of education but serve as instruments of reform by raising standards for all students.

Given our past experience with high stakes assessment and the current state of our knowledge about the construct validity of performance assessments, this anticipation is optimistic at best—the assumptions about the quality of information and the consequences of the assessment must be carefully evaluated. Dunbar, Koretz, and Hoover (1991) note that "the nation stands poised on the brink of yet another wave of test-based reform, and again we appear prepared to undertake it without sufficient quality control" (p. 302). As research into the consequences of high stakes assessment suggests, choices made in designing assessment systems not only impact the nature of teaching and learning (in both intended and unintended ways) but also the nature of the discourse about the purposes and processes of education. Some have begun to raise questions about whether assessments should be used to influence teaching and learning in the instrumental ways proposed in much of the current rhetoric (e.g., Bryk & Hermanson, 1993; Darling-Hammond & Snyder, 1992).

Here I sketch two alternative models of assessment that reflect very different understandings of the role assessment can play in informing stakeholders and promoting educational reform. Each provides information about individual students to be used in decisions about readiness for graduation and information about the educational system to be used for public accountability. Although the particular models are hypothetical, the features are typical of those to be found in actual assessment practice or development. Collectively, the features represent ambitious programs of assessment. I intend both as examples of assessment practice that thought-

ful people find sound. The comparisons which I want to highlight are not between good and bad assessment practice, but rather between different views of the ways in which assessment can work to promote both quality and equality in education. Features of the first model are probably more widely represented in current assessment practice, although I must acknowledge a strong preference for the second model and will argue that point in my comparative comments. The comparison should help to highlight the assumptions and values underlying each model and, I hope, to promote thoughtful discussion of alternatives.

School A in State A

In schools in one state, all students are required to take and pass a statewide writing proficiency test (along with other subject area tests) in order to receive a state-endorsed high school diploma. The test consists of two writing tasks and a multiple choice test of standard written English conventions. The writing tasks encourage analytical thinking by presenting students with an issue and background information reflecting multiple perspectives to use in reaching and justifying a decision. Each task is administered over two class periods, thus providing some opportunity for planning and revision. The written essays are each scored independently and anonymously by two readers who have been carefully trained to use a holistic scoring guide developed by the state. The four scores (two readers for each of two prompts) are combined with the multiple choice score to form an overall composite score. The pass/fail decision is based upon a comparison of the composite score to a cut score that has been previously set by a representative committee of teachers and other stakeholders. The score, pass/fail decision, and information about how to interpret the score (i.e., what kind of performance is typical of students receiving that score), is shared with students, teachers, and parents. Students who did not pass the test have multiple opportunities to retake it following additional instruction and practice.

The test was developed in extensive consultation with teachers, parents, policy makers, and other stakeholders. Careful analysis of the existing curriculum, publication of relevant instructional materials, and widespread inservice opportunities for all teachers suggest that all students have had opportunity to learn the capabilities assessed. Extensive pilot testing and ongoing data collecting show that evidence of construct validity, including reliability, is consistent with sound professional practice for both individual and system level purposes.

To provide system level information for public accountability and policy decisions, the individual scores are aggregated and reported publicly for the state as a whole and for each school and district. Schools with unacceptable passing rates for three years in a row are referred to the state's school

improvement program for consultation and assistance. To provide evidence relevant to equity concerns, aggregate information is reported separately for groups distinguished by gender, race and ethnicity, school location, and socioeconomic status of the community. In addition to these and other measures of student achievement, the state requires presentation of indicators about the school context, including resources and learning opportunities, and additional outcome information, such as drop-out and attendance rates, for policy makers to use in interpreting results. (See examples of the features reflected in State A described in Baker, 1989; Haertel, 1989; OTA, 1992, and Linn, 1993).

School B in State B

In the other state, schools and/or districts are required to develop their own plans for certifying students for graduation and documenting the validity of those decisions. Faculty in some schools have chosen a process that looks something like a dissertation exam. Students prepare a portfolio of their work consisting of a reflection on their development and accomplishments in light of their personal learning goals, work exhibits that demonstrate their capabilities with respect to school and state curriculum standards, and letters of recommendation from those who know the students' work well in and out of school. Plans for preparing the portfolio begin early in high school and preparation of the work constitutes one important part of the curriculum. At least one project, for instance, is expected to show evidence of purposeful, knowledge based, disciplined inquiry. Students complete these projects over extended periods of time, drawing on existing resources, gathering new evidence, sharing work in progress with teachers and peers for critique, and preparing a final exhibit of their work. The portfolio is shared with a committee consisting of the students' teachers, a teacher who has not worked closely with the student, another student, and a member of the community. The committee meets to discuss the work with the student and then to debate its merits in light of school and state curriculum standards. The committee chair prepares a summary of the discussion to document the rationale for the initial decision about the students' readiness for graduation, and, where the initial decision is negative, offers suggestions for additional work. The decision, rationale, and suggestions are shared with the students and their parents. Students have multiple opportunities to rework and resubmit their portfolios to the committee for review. All portfolios and accompanying decisions are audited and confirmed by a school administrator and an affirmative action officer. Students who disagree with the decisions have the right to appeal. Periodically, committees at the district and state levels audit samples of portfolios, decisions, and rationales to assure that the school-level committees are using appropriate procedures and standards. (See, for instance, examples

described in Berlak et al., 1992; Darling-Hammond & Snyder, 1992; Perrone, 1991). To provide school-level information for public accountability, a pilot project is underway that draws on qualitative research methods to develop a category scheme for characterizing the nature, substance, and quality of the work contained in the portfolios. The plan is to randomly sample a small percentage of portfolios and to use the category scheme developed to present a school level portrait, accompanied by samples of students' work. (See similar examples in Applebee, 1981, 1984; Moss, Gere, Clark, & Muchmore, 1993; NAEP, 1990).

To provide information at the state level, a series of both brief and extended performance assessments are administered to students in grades five, eight, and eleven. The state uses matrix sampling—any one student takes only a few of the tasks (one extended task and a few briefer tasks), but a wide variety of tasks are administered to varying samples of students, providing a rich and varied portrait of educational achievement in the state. Results are released at the state and district level, but not at the school or individual level. As with State A, information to inform questions of equity is provided separately for groups distinguished by gender, race and ethnicity, school location, and socioeconomic status of community. Again, as with state A, the state also provides information on a rich variety of indicators of the learning context and resources to use in interpreting results. (see, e.g., any of the recent NAEP Report Cards, 1990.)

Learning from the Comparison

A comparison of these examples highlights a number of issues that should be considered in the design of any high stakes assessment system—issues related to the nature and quality of the information, the way in which it is used, and the impact of the system on teaching, learning, and discourse about educational reform. Both schools-within-states are using assessments to serve two purposes: to provide system-level information for public accountability and to provide individual-level information about students' progress toward state and school curriculum standards to inform decisions about readiness for graduation. A major difference between the two examples involves the level of coordination between assessments at the system and individual levels. In State A/School A, a single assessment is being used to serve both individual- and system-level purposes. This results in census testing where all students in the state are given the same or comparable assessments. Authority for determining the nature of the assessment serving both individual- and system-level purposes is located centrally at the state level. Although state assessment developers were inclusive and democratic in selecting the committee to design and review the assessments, once developed, they are centrally controlled. In State B/School B, the assessments for serving individual and system levels pur-

poses are distinct. For the system-level purposes, matrix sampling is used and only a relatively small proportion of students are tested. At the individual level, authority for determining the nature of the assessments rests with the faculty (and students) in the school, although the faculty is accountable to the district and state, indirectly, through periodic audits of their assessment procedures. Only authority for determining the system level assessment is centrally located. Although all students in the state are expected to document their readiness for graduation through assessments designed by the school, only samples of students respond to the state designed assessment. Existing evidence about the consequences of high stakes testing allows us to anticipate a number of implications for the students and teachers and other stakeholders in these states.

At the *state level*, the assessments are similar in many respects—both are standardized performance assessments. The major differences concern who takes the test—all students versus a sample—and at what level of disaggregation the results are reported—for students, schools, and districts versus for districts only. State A, which tests every student preparing for graduation, offers information about a considerably narrower array of educational outcomes than State B. Given a fixed budget, there is a strong inverse relationship between the number of students that can be tested and the complexity, breadth and depth, of the information that can be provided. In the context of public accountability or program evaluation, it is not necessary to assess every student to obtain valid estimates of performance for the system. Assessing every individual limits the scope and depth of outcomes that can be assessed which in turn limits the information available. A far broader range of educational outcomes, including work that reflects extended discourse, can be feasibly tested if careful sampling is conducted.

The stakes associated with performance on the state assessment in State A/School A are considerably higher than those for State B/School B. Even if certification for graduation and consignment to the school improvement program were not directly tied to performance, whenever scores are released at the level of schools and individual students, there is considerable pressure to improve performance on the test. Teachers and students in low-scoring schools are likely to spend a large amount of their instructional time preparing for this test. While this may have a positive impact in schools where teachers have not previously encouraged the activities required by the test, it may divert attention from additional, equally important, but less easily measured, outcomes. Moreover, to the extent that instruction focuses on the form and content of the assessments, the tests become less valid as indicators of the broader capabilities they are intended to tap, thus undermining the inferences we can draw about the overall quality of education.

One drawback for School B/State B is the lack of school-level information from the state-wide assessment. Although School B is developing a strategy for providing a school-wide portrait, there is no comparative information for public accountability. Moreover, because scores are not released at the school (or student) level, there is some concern that students and teachers will not be motivated to do their best work and that the results will underestimate students' capabilities. The choice not to release scores below the district level involves trade-offs among competing concerns. To provide valid estimates of school-level achievement, the number of students sampled would need to be larger—thus reducing the breadth and depth of outcomes that could be feasibly assessed at the state level (within a fixed budget). Further, to provide information at the school level would have raised the stakes of the state level assessment, making it far more likely that it would narrow teaching and learning to focus on the form and content of the tests in ways that policy makers in State B did not intend. For those relying on assessments to directly alter educational practice, this may be a drawback, but, as I'll elaborate below, policy makers and educators in State B are relying on a far less instrumental model of educational reform.

Both states appropriately provide an array of additional indicators about students' background and the school context to use in interpreting the results. Such information is crucial in guarding against misleading conclusions. For instance, changes in assessment scores from year to year may simply reflect changes in the student population (dropouts or transfers, for instance) rather than changes in the capabilities of students. Differences in assessment scores across ethnic groups may reflect differences in socioeconomic status and resources of the communities in which they live. Differences in assessment scores from school to school may reflect differences in learning opportunities such as the qualification of teachers or the number of advanced course offerings. As Darling-Hammond & Ascher (1991) note, "comparisons of test scores that ignore these factors hold little promise of directing policy makers' attention to the real sources of the problem, so that it can be rectified (p. 16)." [See articles on the design of indicator systems by Bryk & Hermanson (1993), Darling–Hammond & Ascher (1991), and Oakes (1989).]

At the *individual level*, both schools are using assessment information to monitor students' progress toward state and/or school standards and to inform decisions about readiness for graduation. At this level, the assessment systems are radically different, however. In both cases, it is appropriate to raise questions about the soundness of the inferences about students' capabilities (construct validity) as well as about the consequences of using the assessment procedures (consequential validity). Elsewhere, I have addressed the construct validity issues associated with these two

different approaches to assessment (Moss, 1994; Moss, et al., 1992): in State A construct validity is evaluated through traditional psychometric procedures; in State B, construct validity can be evaluated using procedures that grow out of interpretive research traditions. [See also Johnston (1989, 1992) and Hipps (1992) for discussions of the validity issues based on interpretive research methods.] Here, the comparisons I raise focus on the consequences of using the two different procedures, considering the models of teaching and learning and the means of fostering educational reform implicit in each approach.

A comparison of the two approaches in terms of the model of intellectual work that they present for students and teachers highlights the tension between the efficiency, reliability, and comparability permitted by standardization and the collaborative, purposeful work enabled when authority is shared by teachers and students. In school A, all students must take the same state administered test, carefully designed so that all have had the opportunity to learn the necessary capabilities; whereas in school B, each student develops projects, in consultation with faculty, that both suit their own interests and show evidence of having met school and state standards. This represents a different perspective on fairness and on the authority allocated to students in making assessment decisions—in one case allowing students to choose the products that best represent their strengths and interests, and in the other case, presenting all students with the same task after ensuring, to the extent possible, that they have had the opportunity to learn the necessary skills and knowledge.

In school A, the assessments are scored anonymously by readers from outside the school working independently from detailed scoring guides; the decision about readiness for graduation is then made, algorithmically, by aggregating scores across readers and performances and comparing the aggregated score to a previously determined pass/fail cut score. In school B, the assessments are evaluated by teachers who know the students' work well, in dialogue with one another and with those who know the student less well, debating the merits of the performance in terms of school and state curriculum standards. Teachers in School A, along with students and their parents, become consumers of the interpretations constructed by others. Teachers in School B, in critical dialogue with one another, and with students and parents, construct, critique, and revise interpretations about students' capabilities based on available evidence. Again, the approaches reflect a different view of fairness to students and of authority allocated to teachers—one based on anonymity and multiple independent readings; the other based on in-depth knowledge and critical dialogue. [See Moss (1994) for a more extended discussion of this issue.]

As with the state-level assessments, there are trade-offs—possibilities provided in one model that are not possible in the other. One obvious

drawback for School B/State B is that students, teachers, and parents do not receive information that allows normative comparisons with other students in the state. However, to provide such information would again increase the power of the assessment to focus the curriculum toward the form and content of the test—an outcome that policy makers in State B are seeking to avoid.

Perhaps the most controversial feature of State A's system is the direct connection between assessment results and decisions about readiness for graduation. From one perspective, this can be viewed as a means of ensuring that all students are being held accountable to the same standards—thus enhancing both equity and excellence. However, as Haertel (1989) notes in his paper for the National Commission on Testing and Public Policy, "simplistic policies, where action is triggered by scores above or below a cutting point on a single test...are contrary to the consensus of professional practice in testing." (p. 32). Although students, appropriately, have multiple opportunities to take the test, they are provided with one means through which they must demonstrate their competence. Moreover, as schools provide remedial instruction to prepare students who failed for a second (or third) try, the focus of their education is likely to become considerably more narrow than that of their higher scoring peers, thus decreasing equity in terms of access to knowledge. Some may argue that the skills assessed *should* be a major focus for instruction because they are prerequisite to more complex learning opportunities—however, these are risky assumptions, often naively made, that require careful empirical support. Past experience with similar assumptions about so-called basic skills have been seriously challenged by empirical evidence (e.g., Resnick & Resnick, 1992), and it is crucial that any such assumption be carefully evaluated.

Moreover, as Darling-Hammond and colleagues (Darling-Hammond, 1989; Darling-Hammond & Ascher, 1991; Darling-Hammond & Snyder, 1992) argue, it is important to consider whether such judgments are better made locally, grounded in the contextualized and professional judgments of teachers, or at a distance, grounded in the policies and regulations of states or districts. Darling-Hammond and colleagues suggest that matters of equity, such as the allocation of resources and guarantees of equal access, where competing interests exist, should be regulated through bureaucratic mechanisms to protect the needs of all concerned. Issues of productivity, however, such as improving student and school achievement, may be best handled through professional judgment, because needs vary and decisions are best made by those most knowledgeable about the students and the school context. As they note, decisions about students' needs are often far too complex to be prescribed from afar, and teachers' work should be structured to ensure that they are able to make responsible,

knowledge-based decisions. And, as I've argued elsewhere, the validity and fairness of such decisions can be evaluated through critical, evidence-based dialogue, audit, and appeal—similar to the way decisions are made and warranted in the law.

Ultimately, the assessment policy of these two states reflects two different visions of how educational reform is best fostered. State A has instituted a policy that, whether intended or not, promotes change by attempting to control teaching and learning through strong incentives associated with externally imposed assessments. State B has instituted a policy that provides system level information at the state and district level while permitting autonomy at the school level—an autonomy that can encourage purposeful and collaborative activity by teachers and students. Bryk and Hermanson (1993) offer a useful distinction between two different views of the ways in which indicators enter and influence the discourse and practice of educational reform: an "instrumental use" model and an "enlightenment" model. In "the instrumental use" model, the goals are: to develop a comprehensive set of outcome measures; to examine the relationship between these outcomes and indicators of school resources and processes; and, based upon that generalized knowledge, to control schools through allocation of resources, rewards and sanctions, and regulation so as to maximize performance on the outcomes. As they note, the instrumental use model characterizes much of the current rhetoric about the potential of indicators to improve schools. In criticizing this conceptualization, they argue first that there are many valued outcomes for which available measures do not exist. As our past experience suggests, any model which attempts to maximize measurable outcomes is likely to result in a variety of unintended, possibly undesirable, effects, including the undermining of progress in areas not addressed. More fundamentally, the instrumental use model, with its reliance on generalizations about the relationship between processes and outcomes, underrepresents the complexity of schools. While "external policy-making and administrative action shape schools' structure and function" (p. 453), the "behavior, attitudes, and beliefs of actors inside the school—professional staff, parents, and students—influence its operations" (p. 453). "Schools are places where personal meaning and human intentionality matter." (p. 457) An "enlightenment model" of information use reflects a view of schools where interaction among individuals is fundamental and reform requires "changing the values and tacit understandings that ground these interactions" (p. 453). From this perspective, the goal of an indicator system is not to manipulate such interactions through external controls, but rather to "enrich and encourage sustained conversation about education, its institutions, and its processes in order ultimately to improve them" (p. 467).

CONCLUSION

Recognizing that not all valued educational outcomes can be measured with standardized assessments and that evaluation impacts teaching and learning, we need to consider how to design assessment systems that document and promote a wider range of valued educational goals. While we are becoming increasingly knowledgeable about how to design and evaluate standardized forms of performance assessment, we are considerably less knowledgeable about how to design and evaluate nonstandardized assessments and about how to incorporate them into our on-going assessment practices. To these ends, measurement professionals need to look beyond the boundaries of psychometrics to support the validity of nonstandardized forms of assessment, where appropriate, that honor the purposes of students and the contextualized judgments of teachers. Teachers need to assume more responsibility for accounting for their own practice through collaborative inquiry and ongoing peer review, so that their voices are not overshadowed by externally imposed assessments; and administrators need to provide them with the time and resources to do so. Parents, along with other members of the community and the journalists who inform them, need to question whether reliance on easily consumed summary statistics is counter-productive and to consider other means of becoming informed about the quality of education students are receiving. The point is not to overturn the use of standardized assessments (performance based or multiple choice), but to consider carefully the role that they should serve in conjunction with other forms of assessments.

As the existing evidence indicates, assessment systems provide more than indicators of students' achievements: they provide potent and value-laden models of the purposes and processes of school, of the appropriate roles for teachers, students and other stakeholders in the discourse of teaching and learning, and of the means through which educational reform is best fostered. Often in the past, policy makers have been too quick to implement assessment systems without adequate attention to the potential and actual consequences of their actions. There are no simple resolutions to the tensions I've raised. Those who implement assessment policy need to carefully evaluate both the quality of the information and the intended and unintended consequences of using assessments. Before any assessment is operationalized, policy makers should become informed about the existing research on the consequences of various assessment choices, compare alternative approaches to assessment in light of the differential consequences they might foster, and explicitly evaluate the vision of education implied in those consequences. After the system is implemented, policymakers should hold themselves accountable through ongoing monitoring of the consequences of their actions. Good intentions may not materialize,

or worse, may result in untenable outcomes, such as when remedial instruction for lower scoring students results in differential access to knowledge. Nothing should be taken for granted. As Bryk and Hermanson (1993) note, "more information is not always better.... The ultimate long-term test of this system is not whether we are better informed but whether we act more prudently. In the shorter term, the best 'test' may be found in the answer to the question "Is our public discourse enriched (or impoverished) by this new information?'" (p. 476).

REFERENCES

AERA, APA, & NCME. (1985). *Standards for educational and psychological testing.* Washington, DC: Authors.

Applebee, A.N. (1981). *Writing in the secondary school: English and the content areas.* Urbana, IL: National Council of Teachers of English.

Applebee, A.N. (1984). *Contexts for learning to write: Studies of secondary school instruction.* Urbana, IL: National Council of Teachers of English.

Baker, E.L. (1989). Mandated tests: Educational reform or quality indicator? In B.R. Gifford (Ed.), *Test policy and test performance: Education, language, and culture* (pp. 3–23). Boston: Kluwer.

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

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

Bryk, A.S., & Hermanson, K.M. (1993). Educational indicator systems: Observations on their structure, interpretation, and use. *Review of Research in Education, 19,* 451–484.

Camp, R. (1992a). The place of portfolios in our changing views of writing assessment. In R. Bennett & W. Ward (Eds.), *Construction versus choice in cognitive measurement.* Hillsdale, NJ: Erlbaum.

Camp, R. (1992b). Assessment in the context of schools and school change. In H.H. Marshall (Ed.), *Supporting student learning* (pp. 241–263). Norwood, NJ: Ablex.

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

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

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

Crooks, T.J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 85,* 438–481.

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

Darling-Hammond, L., & Ascher, C. (1991). *Accountability in urban schools.* New York: National Center for Restructuring Education, Schools, and Teaching, Teachers College, Columbia and ERIC Clearinghouse on Urban Education.

Darling-Hammond, L., & Snyder, J. (1992). Reframing accountability: Creating learner-centered schools. In A. Lieberman (Ed.), *The Changing Contexts of Teaching* (Ninety-first Yearbook of the National Society for the Study of Education), (pp. 11–36) Chicago: University of Chicago Press.

Dorr-Bremme, D.W., & Herman, J.L. (1986). *Assessing student achievement: A profile of classroom practices.* Los Angeles: University of California, Center for the Study of Evaluation.

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

Ellwein, M.C., Glass, G.V., & Smith, M.L. (1988). Standards of competence: Propositions on the nature of testing reforms. *Educational Researcher, 17* (8) 4–9.

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.), (pp. 105–146). Washington, DC: The American Council on Education and the National Council on Measurement in Education.

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

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

Haertel, E. (1989). Student achievement tests as tools of educational policy: Practices and consequences. In B.R. Gifford (Ed.), *Test policy and test performance: Education, language, and culture* (pp. 25–50). Boston: Kluwer.

Herman, J.L., & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational measurement: Issues and practice, 12*, 20–25, 41–42.

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

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*, 237–264.

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

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

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

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1992). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program: Interim report* (Center for the Study of Evaluation, CSE Tech. Rep. No. 355). Santa Monica, CA: Rand Institute on Education and Training, National Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D., Stecher, B., & Deibert, E. (1992). *The Vermont Portfolio Assessment Program: Interim report on implementation and impact, 1991-92 School Year* (Center for the Study of Evaluation, CSE Tech. Rep. No. 350). Santa Monica, CA: Rand Institute on Education and Training, National Center for Research on Evaluation, Standards, and Student Testing.

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

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

Madaus, G.F. (1988). Testing and the curriculum: From compliant servant to dictatorial master. In L. Taylor (Ed.) (87th edition). *NSSE Yearbook* (pp. 83–121). Chicago: National Society for the Study of Education.

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

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

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

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

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

Moss, P.A., Beck, J.S., Ebbs, C., Herter, R., Matson, B., Muchmore, J., Steele, D., & Taylor, C. (1992). Portfolios, accountability, and an interpretive approach to validity. *Educational Measurement: Issues and Practice, 11*, 12–21.

Moss, P.A., Gere, A.R., Clark, C.T., & Muchmore, J.A. (1993). *Collaborative inquiry, contextualized assessment, and accountability in urban classrooms: A proposal to the Field Initiated Grants Program.* Unpublished manuscript, University of Michigan.

Moss, P.A., & Herter, R.J. (1993). Assessment, accountability, and authority in urban schools. *The Long Term View, 1*, 68–75.

National Assessment of Educational Progress (NAEP). (1990). *Exploring new methods for collecting students' school-based writing.* Washington, DC: U.S. Department of Education.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America.* Chestnut Hill, MA: Author.

Neill, D.M., & Medina, N.J. (1989, May). Standardized testing: Harmful to educational health. *Phi Delta Kappan*, pp. 688–697.

Newmann, F.M. (1990). Higher order thinking in teaching social studies: A rationale for the assessment of classroom thoughtfulness. *Journal of Curriculum Studies, 22*(1), 41–56.

Nystrand, M., Cohen, A.S., & Dowling, N.M. (1993). Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment, 1*(1), 53–70.

Oakes, J. (1989). What educational indicators? The case of assessing the school context. *Educational Evaluation and Policy Analysis, 11* (2), 181–199.

Oakes, J., Gamoran, A., & Page, R.N. (1992). Curriculum, differentiation: Opportunities, outcomes, and meanings. In P.W. Jackson (Ed.), *Handbook of research on curriculum: A project of the American Educational Research Association* (pp. 570–608). New York: Macmillan Publishing Company.

Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (Technical Report, OTA—SET-520). Washington, DC: Congress of the United States.

Perrone, V. (Ed.). (1991). *Expanding student assessment.* Washington, DC: Association for Supervision and Curriculum Development.

Resnick, L.B., & Resnick, D. (1992). Assessing the thinking curriculum: New tools for educational reform. In. B.R. Gifford & M.C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction (pp. 37–75).* Boston: Kluwer.

Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assessment of writing.* Norwood, NJ: Ablex.

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30* (3), 215–232.

Shepard, L.A. (1992). What policy makers who mandate tests should know about the new psychology of intellectual ability and learning. In. B.R. Gifford & M.C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 301–328). Boston: Kluwer.

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

Smith, M.L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20*, 8–11.

Tavalin, F. (1993). Vermont writing portfolios. In M.A. Smith & M. Ylvisaker (Eds.), *Teachers voices: Portfolios in the classroom.* Berkeley: National Writing Project.

U.S. Department of Education. (1988, 1989). *Digest of Educational Statistics.* Author: Washington, DC.

Wiley, D.E., & Haertel, E.H. (in press). Extended assessment tasks: Purposes, definitions, scoring, and accuracy. In R. Mitchell & M. Kane (Eds.), *Implementing performance assessment: Promises, problems, and challenges*. Washington, DC: Pelavin Associates.

Willinsky, J. (1990). *The new literacy: Redefining reading and writing in the schools*. New York: Routledge.

Wolf, D., Bixby, J., Glenn, J., III, & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17*, 31–74.