Part Three

Constructive Tests for Accountability

Chapter 6

Portfolio Assessment and Education Reform

Brian Stecher

Educational policy makers in a substantial number of states are looking toward assessment to accomplish the dual goals of increasing educational accountability and changing instructional practice. Portfolios have gained favor with many of these individuals because of the belief that portfolios better model the kinds of activities students should engage in while providing scores that are valid for accountability purposes. If you believe that "what you test is what you get," then portfolios may be the form of testing that provides the optimum benefit.[1]

This chapter reviews the evidence on the effectiveness of large-scale portfolio assessment in the United States, including state assessment systems in Vermont and Kentucky, portfolio experiments in Pittsburgh, Pennsylvania, and California, and the experiences of the National Assessment of Educational Progress (NAEP). The evidence indicates that portfolio assessments are relatively weak as tools for educational accountability (in comparison with other assessment methods), but they are relatively strong in influencing educational practice (again, in comparison

with other assessment methods). Even here their promise is limited because they are also costly, both in terms of the cost of scoring and in terms of teacher and student time.

We begin with a discussion of the definition of portfolio assessment and the purposes this form of assessment might serve. Then we review the research on portfolio assessments, including their technical quality, their effects on classroom practices, and the burden/costs they place on students and teachers.

Portfolio Assessment: What and Why

Portfolio assessment of academic subjects draws its inspiration from the world of art. An artist's portfolio contains a purposeful collection of his or her own work. Such a portfolio is highly personal; it can include fully realized images, preliminary sketches, multiple versions of the same piece, and so on. Its contents reveal the skills of the artist as well as the choices the artist makes in assembling the work.

An academic assessment portfolio is a collection of student work that reflects the skills the student has mastered in a particular subject domain. However, as in the case of the artist, the portfolio format permits considerable variation in emphasis. Portfolio assessment guidelines can be designed to capture different aspects of student work. For example, portfolio assessments can be designed to emphasize development of expertise over time by requiring students to collect drafts, revisions, and final work at the beginning, middle, and end of the year. Alternatively, portfolio assessments can be designed to document optimum performance. Vermont students were required to collect five to seven of their "best pieces" in their mathematics portfolios.[2] A portfolio assessment could also be designed to emphasize breadth of understanding. In Kentucky, students were instructed to include particular types of writing in their writing assessment portfolios, including a poem, a persuasive letter, and so on.[3] Thus, the notion of a portfolio assessment is quite broad, and without some further stipulation, portfolio assessments may differ from place to place.

Differences may also arise because portfolio assessments can serve multiple purposes, and the choice of purpose will affect the way the portfolio assessment is structured and implemented. Educators cite at least four purposes for using portfolio assessments. First, portfolios can encourage student reflection and self-evaluation. Reviewing a portfolio of his or her own work can make a student more self-aware, can build a student's understanding of the cumulative nature of learning and the interrelationships among the skills he or she has acquired, and can enhance a student's ability to evaluate his or her own work. Second, portfolios can be used to help teachers monitor student learning, diagnose their strengths and weaknesses, and plan better instruction. By focusing the teacher's attention on student work rather than test scores, the portfolio permits more refined judgments of skill acquisition and contributes to more thoughtful planning for remediation or enrichment. Third, portfolios can encourage curriculum change. Administrators may choose to mandate portfolio assessments to encourage teachers to change the nature of instruction. For example, the implementation of a writing portfolio assessment will force teachers to spend time on extended writing assignments; mathematics portfolio assessments often necessitate more classroom emphasis on mathematical problem solving. Finally, portfolio assessments can be used as a basis for systemwide accountability. Many educators believe portfolio assessments provide more valid information about important student outcomes than do multiple-choice tests. They believe that portfolios focus attention on complex, fully realized products of student activity, whereas standardized, on-demand, multiple-choice tests focus attention on separable facts and disaggregated procedures.

With all this variation in structure and purpose, one must be cautious in making statements about portfolio assessments in general. Not only may portfolio assessments differ substantially one from another, but small differences in implementation can have large differences in consequences. For example, J. R. Novak and colleagues found that small differences in scoring rubrics affected teachers' understanding of narrative writing.[4]

Fortunately, those jurisdictions whose portfolio assessment systems have been studied most thoroughly have implemented systems that share many common features. With slight variations, the portfolio assessments in these locations have the following characteristics:

- Constructed, not selected products. The portfolio contains work produced by students in response to classroom assignments, including such things as written essays, drawings, graphical representations, and so on. Most of the contents are produced by students working alone, but some clearly identified collaboration with other students may be allowed.
- Limited number of pieces. The portfolio contains a small number of examples of student work (roughly three to seven pieces) rather than a complete compendium of work for the year.
- "Embedded" in instruction. The work collected in a student's portfolio has been produced as part of ongoing classroom activities. One consequence of the embedded nature of portfolio assessments is that the contents of the portfolios can vary from teacher to teacher because assignments vary.
- Student choice. Each student selects the pieces to include in his or her portfolio (with varying degrees of input from the teacher). As a result, the pieces that are included in the portfolio will vary from student to student within a given class.
- Cumulative. Each portfolio is accumulated over an extended period of time, and is not created "on demand" like a standardized test.
- Introductory essay. The portfolio contains a description of the contents or a reflective essay in which the student explains his or her choices.
- Scoring system. There is a more or less objective procedure for reviewing the contents of each portfolio and assigning one or more scores to the student.

- Dual purposes. In the sites that have been studied most extensively, portfolios were implemented both to influence instruction and to provide valid scores for accountability purposes.

Evidence About Portfolio Assessment

The bulk of the research on portfolio assessment was conducted in a handful of jurisdictions. These programs and research efforts are summarized below.

The Vermont Portfolio Assessment Program

The Vermont Portfolio Assessment program was implemented in 1990–91 as the first centralized assessment system in the state's history.[5] As such, it received considerable attention within Vermont. Because of its novel use of portfolios it also received considerable attention nationwide. The program began on a pilot basis in 1990–91, and it was made operational the following year. The system assessed each student in fourth and eighth grade in two subjects, mathematics and writing.

The most important elements of the system were portfolios of student work in mathematics and writing. The mathematics portfolio consisted of five to seven "best pieces" selected by the student from all the work done during the year. Mathematics portfolios were scored on seven dimensions, four reflecting aspects of mathematical problem solving (understanding the task, approaches/procedures, decisions along the way, and outcomes of activities) and three reflecting aspects of mathematical communication (language, representation, and presentation). Each dimension was scored on its own four-point scale.

In writing, students selected a single "best piece" and included other writing of specified types, which were graded as a set. Writing portfolios were scored on five dimensions (purpose, organization, details, voice/tone, and usage/mechanics/grammar). Again, each dimension had its own four-point scale.

The goal of the portfolio assessment program was to report dimensional-level scores that would permit comparisons among schools. A random sample of portfolios from each class was sent

for central scoring by teachers. The developers hoped that school-level dimensional scores derived from this sample of students would be valid for school-level accountability.

In addition to assessment portfolios, the Vermont system also included a common test, called the Uniform Test, which was taken by all students. The Uniform Test in mathematics included both multiple-choice and constructed-response components; the Uniform Test in writing consisted of a single writing prompt. Initially, the main purpose of this test was to validate the portfolio scores. Over time, the program has been revised to place greater emphasis on common components. As of 2000, the portfolios are still part of the formal assessment system in Vermont but their role relative to other components has diminished.

Most of the published research on the Vermont program was conducted between 1991 and 1994 by researchers from RAND.[6] These studies examined the technical quality of scores from the mathematics and writing portfolios. However, the researchers focused their investigation of changes in classroom practices on the mathematics portfolios, which were the most innovative part of the system.

Kentucky Instructional Results Information System (KIRIS)

In 1990, the Kentucky legislature enacted the Kentucky Educational Reform Act (KERA) in response to a court mandate to reform the education system in Kentucky. KERA was a comprehensive reform that changed school finance, teacher professional development, the organization of primary grade schooling, and the statewide curriculum. A prominent feature of KERA was a new accountability system for schools based primarily on performance-based assessments rather than multiple-choice tests. KIRIS was developed to comply with the assessment requirements of KERA. The system was initiated in 1990 and continued in operation through the 1997–98 school year.[7] In 1998–99, the system underwent a substantial redesign, and it is now known as the Commonwealth Accountability Testing System (CATS). The writing portfolios were retained, but not the mathematics portfolios.

The KIRIS testing program had a number of different components, including multiple-choice items, constructed-response items, performance events (with a collaborative component), on-demand writing prompts, and portfolios in writing and mathematics. The program covered seven subject areas, including mathematics, science, social studies, writing, reading, arts and humanities, and practical living/vocational studies. Initially, testing was done in grades four, eight, and twelve in all subjects, but the burden on these students and teachers became too great. After about four years, the elementary testing was divided between grades four and five and the middle-school testing between grades seven and eight. Later changes further divided high school testing among three grade levels.

Students were classified into four performance levels in each subject based on their scores on the relevant assessments. The levels were called Novice, Apprentice, Proficient, and Distinguished. Each school received an overall KIRIS accountability score based on the percentage of students achieving the Proficient level of performance in each subject. The accountability formula also gave credit for performance on a set of noncognitive indicators, including attendance and dropout rate (at the secondary level). Schools were expected to improve their performance annually with an eventual goal equivalent to all students reaching the Proficient level. Cash rewards were given to schools for making large gains. Schools that scored poorly and failed to make improvement were assigned Distinguished Educators to work with them to improve their scores. The rewards system operated in two-year cycles, so schools' performance in one biennium was compared with their performance in the previous biennium.

The structure of the portfolios in Kentucky was similar to that in Vermont. Students compiled their assessment portfolios by selecting pieces from a working portfolio they collected all year. Assessment portfolios in writing consisted of six pieces, including a personal narrative; a poem, play, or piece of fiction; a persuasive or informative piece; a piece from another subject area; a best piece; and a letter about growth as a writer. Portfolios were scored by classroom teachers, who assigned a single performance

level to each student based on the whole portfolio. In assigning the score, raters considered six dimensions of performance in writing (purpose and approach, idea development, organization, sentences, wording, and surface features). Writing portfolio scores were the sole indicator of writing proficiency in KIRIS. Mathematics proficiency was measured by a combination of on-demand testing and portfolios. (The mathematics portfolios were collected every year, but the scores were included in the computation of the schools' accountability index during the second biennium only.)

The Kentucky Department of Education encouraged researchers to study the program and contribute to its improvement. In addition, KIRIS was the subject of much controversy in the state because of the more innovative components. As a result, there was considerable research on the technical quality of scores and the effects of KIRIS on classroom practice.[8]

Pittsburgh, Pennsylvaina

The Pittsburgh School District experimented with writing portfolios in grades six through twelve beginning in 1992. Students compiled working portfolios throughout the year, then selected four pieces to be included in their assessment portfolio at the end of the year. They were supposed to choose an important piece, a satisfying piece, an unsatisfying piece, and a free pick. All drafts as well as the final piece and a written reflection on each piece were to be included in the assessment portfolio. The portfolios also included a table of contents, a writing inventory, and a final reflection.

A stratified random sample of portfolios was selected for scoring by trained teachers. Each portfolio was scored as a whole on three six-point scales whose endpoints were "inadequate performance" and "outstanding performance." The scales reflected accomplishment in writing, use of process and strategies for writing, and growth, development, and engagement as a writer. These rubrics were developed during several years of discussions of writing conducted as part of the Arts PROPEL project in Pittsburgh. Two judges rated each portfolio, and if their scores differed by no

more than a point, their scores were summed to produce the final score. If greater differences occurred, a third judge was used to arbitrate.

The Pittsburgh portfolio assessment was studied by researchers at the Educational Testing Service (ETS).[9]

National Assessment of Educational Progress (NAEP)

The NAEP conducted trial assessments using writing portfolios in fourth and eighth grades in 1990 and 1992. These were exploratory efforts to examine the feasibility of using writing portfolios on a larger scale in future assessments. The results were not reported, and there were no consequences for students or teachers. The 1990 trial was mounted somewhat late, and participation was not as complete as desired. Students contributed a single best piece, which was scored using a six-point scale. In 1992, the NAEP prepared participants better, and teachers collected three pieces from each student. The pieces were classified by genre and scored using six-point, genre-specific rubrics. The process was coordinated and studied by researchers at ETS who were responsible for the NAEP.[10]

California Learning Assessment System (CLAS)

Between 1992 and 1994, researchers from ETS worked with the California Department of Education to develop a portfolio assessment component for the CLAS testing program. The portfolio development process focused on mathematics and language arts, and the project was designed to emphasize performance that was consistent with state standards and curriculum frameworks. As part of the developmental effort, portfolios were collected in language arts and mathematics from a sample of students in grades four and eight. The language arts portfolios were scored on two dimensions: constructing meaning and composing and expressing ideas. The mathematics portfolios were scored on three dimensions: mathematical content, communicating mathematics, and putting mathematics to work. Scores were assigned to the portfolio as a whole, not to individual pieces, and students were classified into four performance levels based on their scores.

The levels were called Beginning, Developing, Accomplished, and Exemplary. There was also a classification for portfolios that had "not enough evidence to judge." The research has been reported by the Educational Testing Service.[11]

Findings

Although these five portfolio assessment programs are different in important ways, the research results paint a fairly consistent picture of the quality of scores derived from portfolios, the effects of portfolio assessments on classroom practices, and the burdens/costs of portfolios (in terms of teacher and student time). Portfolio assessments are relatively weak as accountability tools when compared with other forms of assessment. The scores appear to be less reliable and less valid than scores from standardized, multiple-choice tests. In contrast, portfolio assessments have relatively strong effects on practice. Evidence suggests that portfolio assessments lead to changes in classroom practices in the desired direction. However, this change comes at a relatively high price in terms of student and teacher time. The administrative burden on teachers is particularly high, and it does not appear to lessen over time. These findings are explored in the following sections.

Portfolio Assessment As an Accountability Tool

If portfolio scores are to be used for accountability purposes, it is essential that they have adequate technical quality. A careful analysis of the technical quality of portfolio scores should consider three things: the consistency of ratings of individual pieces (rater reliability), the consistency of student performance across pieces (score reliability), and the interpretability of scores (validity). The research reviewed here suggests that portfolio assessments are weak on all three counts. There are substantial differences between the scores assigned by two raters to a given piece of work. Student performance varies from piece to piece. And as a result of these two inconsistencies, it is difficult to interpret overall scores assigned to student portfolios. In addition,

variations in the difficulty of tasks assigned by different teachers and in the conditions under which students prepare their pieces further clouds the interpretation of scores.

The most comprehensive evidence on reliability comes from Vermont and Kentucky, where portfolios were used on the largest scale for the longest periods of time. In both states, raters showed only moderate agreement on the scores they assigned to individual pieces of student work on the dimensions of interest. Koretz summarized the results of research in Vermont in 1993 (the second year of the portfolio assessment program), and these figures are displayed in Table 6.1.[22] The first row reflects the agreement between two raters on individual pieces from the mathematics portfolio and on the "best piece" or the rest of the writing portfolio. Each piece in the mathematics portfolio was rated on seven dimensions, and each component of the writing portfolio was rated on six dimensions. Table 6.1 displays the average value of the correlation between two raters on each dimension. The results are quite low (0.50 or less). (A correlation of 1.0 means that higher scores from one set of raters are always associated with higher scores from the other set of raters and lower scores are always associated with lower scores. A correlation of 0.50 means that scores from one set of raters predict only 25 percent of variance in scores from the other set of raters.)

The results in the second row indicate the degree to which raters agreed with each other about a student's performance on a dimension after combining ratings across all the pieces in the portfolio. These values are higher than the first row, particularly in mathematics, which may have to do with the number of pieces

TABLE 6.1 Average Inter-reader Correlations on Vermont Portfolio Assessments in Mathematics and Writing, 1993

	Writing	*Mathematics*
One piece, one dimension	.45	.50
All pieces, one dimension	.52	.65
All pieces, all dimensions	.63	.79

(five to seven in math; only "best" and "rest" in reading). Nevertheless, the correlations are still too low to permit reporting of dimension-level scores, which was the goal of the portfolio assessment program. The inability to achieve this goal was a great blow to the program's developers. In the end, they reported only total scores combining all pieces and all dimensions. The third row of the table shows the correlation between raters on total scores. Even these values are relatively low and do not offer great confidence in the accuracy of the overall rating process.

Over time, the consistency of ratings in Vermont improved, particularly in math. Table 6.2 presents similar results for total scores (summed across all pieces and all dimensions) over the four-year period from 1991 to 1995. With training and improvement in rubrics, total math scores reached acceptable levels of reliability, but this was never the case for dimension-level scores, and it was not true in writing.

Table 6.3 contains similar results for writing portfolio assessments conducted in Kentucky and Pittsburgh. In Kentucky, inter-rater correlations on total writing scores across pieces and dimensions were similar to those achieved in later years of the Vermont program. Koretz computed the comparable statistics for Pittsburgh.[13] He reported correlations at the dimension level that were of similar magnitude. Results from the NAEP were similar.

TABLE 6.2 Inter-rater Correlations on Total Scores on Vermont Portfolio Assessments in Mathematics and Writing, 1991–1995

	1991–92	*1992–93*	*1993–94*	*1994–95*
Writing				
Grade 4	.49	.56	.74	.64
Grade 8	.60	.63	.69	.66
Mathematics				
Grade 4	.60	.72	.76	.80
Grade 8	.53	.79	.83	.89

Rater agreement during the second year of the NAEP trial writing assessment was marginally lower than the values reported in Table 6.3. Inter-rater correlations on narrative, informative, persuasive, and overall writing scores in grades four and eight on the NAEP fell in the range of .59 to .68.

An additional source of inconsistency in portfolio assessment scores comes from differences in the performance of students across pieces within a portfolio. This was demonstrated in mathematics (more than writing) in Vermont in 1992. An analysis of dimension-level scores in grades four and eight found a substantial student-by-piece interaction, which means that students performed relatively differently from one piece to the next.[14]

Given the inconsistencies in scoring reported above, it is not surprising that portfolio assessment scores cannot be interpreted in the intended manner. Another way to say this is that the scores lack validity for the intended interpretation. The best way to judge the validity of portfolio assessments is to compare the scores with other measures of similar and dissimilar performance. Scores should converge with measures of similar domains and diverge with measures of different domains. For example, one would expect a new test of math to correlate more highly with another test of math than with a test of social studies. If this does not occur, it calls into question the meaning of the new test score.

There are two instances in which portfolio assessment scores can be analyzed in this manner. In Vermont, researchers compared scores on uniform tests of math and writing with scores on

TABLE 6.3 Mean Inter-reader Correlations on Writing Portfolio Assessments, Kentucky 1993–94 and Pittsburgh 1992

	Total Writing Score (Kentucky)	*Writing Dimensions[a] (Pittsburgh)*
Grade 4	.67	
Grade 8/Middle school	.70	.60 – .67
High school		.71 – .77

a. As reported by Koretz, 1998.

math portfolios, and they did not find the expected pattern of results. Table 6.4 shows the correlations of mathematics portfolio total scores with Uniform Test scores in mathematics and writing (both organization and usage). Surprisingly, the correlations between math portfolios and math Uniform Tests were no greater than correlations with writing Uniform Tests.[15] This raises serious questions about the meaning of scores from the mathematics portfolio assessment.

Similar questions are raised by results from the NAEP writing portfolio trial. The correspondence between scores from the writing portfolios and scores on the on-demand NAEP writing sample was no better than chance.[16] In fact, the correlation between scores was only 0.15.[17] Although these two assessments are trying to tap somewhat different aspects of writing, this level of similarity seems unconvincing. In Kentucky, an expert panel that convened to evaluate KIRIS found inadequate evidence to support the use of the scores for their intended purpose. The convergent and divergent patterns of scores from the various testing components of KIRIS was not convincing.[18]

Finally, it is difficult to use portfolio assessments for comparative purposes when the conditions under which pieces are produced are not standardized. In Vermont, 44 percent of eighth-grade teachers and 65 percent of fourth-grade teachers placed limits on the kind of assistance they provided to students completing pieces; the remainder of teachers did not. One-fourth of the teachers said students' pieces were not revised at all, whereas at the other extreme, 10 percent reported that the average piece was revised three times. Similar discrepancies in the

TABLE 6.4 Average Correlations[a] Between Mathematics Portfolio Total Score and Vermont Uniform Test (UT), 1993

	Writing UT: Organization	*Writing UT: Usage*	*Math UT*
Grade 4	.33	.33	.35
Grade 8	.35	.38	.31

a. Disattenuated for unreliability of raters.

conditions under which portfolio pieces were produced were found in Kentucky.[19]

There are a number of factors that may explain the low reliability of scores from portfolio assessments. First, the uniqueness of each portfolio (which advocates cite as one of the strengths of the approach) forces developers to create very generalized scoring rules. Because each portfolio will contain different pieces of work, scoring rubrics have to be quite generic. As a result, they may not provide enough guidance to raters to insure comparability of ratings. Second, portfolios can include very complex and elaborate assignments (another of their strengths). However, student performance may vary greatly between complex tasks. Such variation reduces the consistency of scores. Third, complex tasks take a long time to complete as well as a long time to score, so most portfolio assessments limit the contents to a relatively few pieces. This exacerbates the problems created by variation in student performance because there are fewer performances to judge.

The low reliability of portfolio assessment scores automatically limits their validity. If the scores themselves are not accurate, they are unlikely to be consistent with other measures. Moreover, most portfolio assessments are implemented as part of efforts to reform curriculum. They are designed to measure constructs that are not well measured by existing tests. Under these circumstances it can be difficult to specify in advance what pattern of relationships among measures is anticipated. If there are no clear expectations to begin with, it can be difficult to determine whether the pattern of results is consistent with expectations.

Portfolio Assessment As a Curriculum Reform Tool

Evidence suggests that portfolio assessments encourage changes in curriculum and instruction. In Vermont and Kentucky, where these changes have been studied most thoroughly, the introduction of portfolio assessment has led to changes that were consistent with the goals of the accompanying reform effort. These reforms, like those in Pittsburgh and California, emphasized "authentic" curriculum (for example, writing with purpose and audience in mind, mathematical problem solving). Most also

emphasized changes in instructional practices to make classroom interactions more "student-centered," that is, giving students more responsibility for structuring and monitoring their own work and encouraging teachers to act more as facilitators. The evidence about changes in curriculum associated with portfolio assessments is the strongest, but there is also evidence about changes in instruction.

The Vermont mathematics portfolio assessment was designed to emphasize problem solving and mathematical communication rather than algorithms and computation, and teachers reported changing their curriculum accordingly. There were widespread increases in the time spent on problem solving and mathematical communication.[20] Approximately three-quarters of teachers said students spent more time making charts, graphs, and diagrams (70 percent), writing reports about mathematics (70 percent), and applying mathematical knowledge to new situations (75 percent). One-half of the teachers in Vermont reported that their classes devoted more time to exploring mathematical patterns.

Researchers in Kentucky found similar curriculum changes in mathematics and language arts, which were consistent with the goals of KERA.[21] In mathematics, teachers reported spending more class time on problem solving and communication and less on number facts. In language arts, teachers indicated that more class time was devoted to writing for a variety of purposes and on analysis of texts and less time to spelling, punctuation, and grammar.

Instructional practices are somewhat harder to measure than curriculum, but researchers reported changes in teaching associated with portfolio assessments in Vermont, Kentucky, Pittsburgh, and California.[22] The reported effects included:

- instructional changes (California).
- increases in the amount of time that learning occurs in pairs or small groups (Vermont).
- more innovative lesson planning (Vermont).
- increases in instruction leading to complex thinking and problem solving (Vermont).

- greater use of open-ended questions (Kentucky).
- increases in student choice of ideas for writing (Kentucky).
- curriculum and instruction changes in writing (Pittsburgh).

Researchers have also identified some negative effects that may be attributable to portfolio assessments in a high-stakes context. Kentucky teachers shifted curriculum in questionable ways in reaction to the grade-specific accountability system used by KIRIS.[23] Many fourth-grade teachers increased the time that students spent studying subjects tested in fourth grade (writing, reading, and science), whereas many fifth-grade teachers increased the time students spent studying subjects tested in fifth grade (mathematics, arts and humanities, social studies, and practical living/vocational studies). This curriculum shift is understandable given the high-stakes testing environment created by KIRIS; however, it is also troubling. Annual changes in the balance among subjects are not part of the Kentucky reform plan, and their long-term impact on student achievement is unknown. Researchers in Vermont reported a subtle form of curriculum narrowing as a result of the scoring rubrics used with the high-stakes portfolio assessment. They found that teachers focused instruction on the aspect of portfolios that scored well rather than the broader domain of knowledge the portfolios were supposed to reflect.[24] They called the phenomenon "rubric-driven instruction," and suggested that in Vermont, rather than "what you test is what you get," they were finding that "what you score is what you get."

The Costs and Burdens of Portfolio Assessment

Portfolio assessments also generate added costs and burdens. This form of assessment is more costly to develop and to operate than standardized tests, and it places greater demands on teachers and students. There are few good estimates of actual costs, but the additional demands placed on teachers have been well documented. These added burdens include additional preparation time, more classroom time for completing tasks and for managing portfolios, and added scoring time. Moreover, these burdens

did not diminish during the first couple of years that portfolio assessment was operational. However, both principals and teachers felt the benefits outweighed the burdens, at least in the early years of the program.

Some of the additional costs and burdens associated with portfolios are easy to quantify, whereas others are quite difficult to measure. The operational costs are borne primarily by the jurisdiction responsible for the assessment. They include the cost of designing the system, specifying the type of student work desired, developing scoring guides, training teachers to understand the assessment, and organizing scoring and reporting. At present there are no comprehensive estimates of the total cost of operating portfolio assessments in any of the jurisdictions that have been studied. However, there are some reports that illuminate the cost of selected components. For example, researchers reported that every teacher in Vermont attended two days of paid professional development each summer for the first two years of the program, but they did not estimate the cost of this training. Kentucky developed a statewide system of professional development centers to support KERA and KIRIS. Although the total cost of these centers was millions of dollars, no one has estimated the cost associated specifically with KIRIS. The Vermont Department of Education paid teachers to come together during the summer to score a random sample of portfolios from across the state. Researchers estimated that the costs associated with this scoring were at least $13 per portfolio, which is more than twice the cost of scoring and reporting services for most standardized multiple-choice tests.[25]

The responsibility for implementing portfolio assessments falls most heavily on teachers. Table 6.5 shows the average number of hours per month that teachers in Vermont and Kentucky devoted to three types of activities in support of the mathematics portfolios.[26] The greatest demands related to preparation, and these ranged from ten to twelve hours per month on average. Teachers participated in professional development workshops to learn about the assessment, and they had to prepare lessons and activities to generate appropriate student work. Vermont teachers

reported that they spent additional preparation time on the following activities (in order of frequency): preparing portfolio lessons, finding appropriate tasks or materials, attending professional development workshops, and discussing portfolios with colleagues.

The portfolio assessment also placed substantial demands on classroom time. Teachers in the two states reported that students spent ten to fourteen hours per month in class working on mathematics portfolio pieces. Vermont teachers reported that classroom time was devoted to completing tasks for the first time, revising tasks, and organizing portfolio materials, in that order. In Kentucky, the classroom time associated with portfolios was devoted to teaching the skills needed to prepare students, doing pieces for the first time, revising/rewriting, and organizing/managing, in that order. Although the bulk of these activities are certainly associated with learning, this still represents a substantial shift in instructional emphasis. Teachers reported that they were taking the time from other instructional activities to devote to portfolio projects. In fact, almost all teachers said it was difficult to cover the curriculum because of the demands of the portfolios. For the most part, teachers reported reducing the time they spent on the mechanical aspects of mathematics, such as computation.

And finally, Table 6.5 shows that teachers spent a great deal of time outside of class scoring student portfolios. In Vermont, researchers reported the average scoring time for a typical month; in Kentucky, scoring was concentrated during a specific period in the spring, and researchers reported total scoring time during this

TABLE 6.5 Average Weekly Teacher Time Devoted to Mathematics Portfolio Activities, Vermont (fourth and eighth grade) and Kentucky (eighth grade)

	Vermont	*Kentucky*
Preparation	12	10
Class time	14	10
Scoring	5	20[a]

[a] Total hours during the scoring period.

period. Teachers in both states felt that scoring was much too time-consuming. It is worth noting that the desire for more accurate scores may lead to even greater demands on time. The Pittsburgh experience suggests that scoring is improved by in-depth, extended, thoughtful discussions to develop shared interpretive frameworks.[27]

Because most of this research was conducted early in the life of the portfolio assessment program, one might expect that demands on teacher and student time would decrease. However, any decrease that did occur during the period investigated by these researchers was quite small. For example, during the first year of the portfolio assessment, 60 percent of the Vermont teachers said they lacked adequate time to prepare. More than two-thirds of the teachers surveyed the next year said the burden had not decreased. Similarly, in the third year of the Vermont program, 80 percent of teachers said scoring was still too time-consuming.

Although the demands of portfolio assessments were great and principals and teachers complained about the amount of time they devoted to the portfolios, on balance, both groups were enthusiastic about the reforms. Researchers characterized the Vermont portfolio assessment as a "worthwhile burden" in the minds of Vermont teachers and principals. In addition, a substantial proportion of Vermont principals said they were going to expand the use of portfolios to other, nontested grades. This is a relatively strong endorsement given their criticism of the additional demands created by the portfolios. What is unclear is how long this endorsement will continue if the portfolios fail to achieve greater reliability and validity and if the burdens do not decline.

Summary

Although the number of jurisdictions using portfolio assessments is small, they have been implemented and studied in enough locations to warrant initial conclusions about their utility. The evidence supports the conclusion that flexible portfolios that

reflect differences in teachers' instructional emphases and students' choice of pieces have not achieved sufficient reliability or validity to be used for the purposes of accountability. The shortcomings derive in large part from the difficulty of developing scoring rubrics that are general enough to apply to widely different pieces, but specific enough to produce agreement among raters. This weakness, coupled with the wide variation in individual performance, leads to scores that do not appear to reflect the constructs the portfolios were designed to measure.

Nevertheless, portfolio assessments have some advantages over other types of assessment. They appear to be strong levers for change in curriculum and instruction. There is ample evidence that portfolio assessments encourage changes in curriculum that are consistent with related reforms—for example, mathematical problem solving and writing for specific audiences. They also promote changes in instruction.

However, these changes come at a price. Portfolio assessments are more expensive to develop and maintain than multiple-choice testing programs. Scoring, in particular, is costly. More important, portfolio assessments impose substantial burdens on teachers, in terms of preparation, classroom activities, and scoring. These burdens do not appear to diminish substantially during the first couple of years of implementation. Perhaps the best role for portfolio assessment is not as an accountability measure, but as a classroom-based assessment tool to help students and teachers improve diagnosis and instruction. This use may maximize the positive aspects of portfolios while minimizing their negative effects.

Notes

1. L. B. Resnick and D. P. Resnick, "Assessing the Thinking Curriculum: New Tools for Educational Reform," in *Future Assessments: Changing Views of Aptitude, Achievement, and Instruction,* ed. B. Gifford and M. C. O'Connor (Boston, Mass.: Kluwer, 1992).
2. Vermont Department of Education, *Looking Beyond 'The Answer': Vermont's Mathematics Portfolio Assessment Program* (Montpelier, Vt.: Vermont Department of Education, 1991).

3. Kentucky Department of Education, *Kentucky Instructional Results Information System: 1991–92 Technical Report* (Frankfort, Ky.: Kentucky Department of Education, 1993). Kentucky Department of Education, *Kentucky Instructional Results Information System: 1992–93 Technical Report* (Frankfort, Ky.: Kentucky Department of Education, 1994).
4. J. R. Novak, J. L. Herman, and M. Gearhart, "Issues in Portfolio Assessment: The Scorability of Narrative Collections" (CSE Technical Report No. 410, Los Angeles, Calif.: CRESST/UCLA, May 1996).
5. Vermont Department of Education, *Looking Beyond 'The Answer'*; R. P. Mills and W. R. Brewer, *Working Together to Show Results: An Approach to School Accountability in Vermont* (Montpelier, Vt.: Vermont Department of Education, October 18/November 10, 1988).
6. D. Koretz, B. Stecher, and E. Deibert, "The Vermont Portfolio Assessment Program: Interim Report on Implementation and Impact, 1991–92 School Year" (CSE Technical Report No. 350, Los Angeles, Calif.: CRESST/UCLA, August 1992). D. Koretz, B. Stecher, S. Klein, D. McCaffrey, and E. Deibert, "Can Portfolios Assess Student Performance and Influence Instruction? The 1991–92 Vermont Experience" (CSE Technical Report No. 371, Los Angeles, Calif.: CRESST/UCLA, December 1993); D. Koretz, B. Stecher, S. Klein, and D. McCaffrey, "The Evolution of a Portfolio Program: The Impact and Quality of the Vermont Program in Its Second Year (1992–93)" (CSE Technical Report No. 385, Los Angeles, Calif.: CRESST/UCLA, July 1994a); D. Koretz, B. Stecher, S. Klein, and D. McCaffrey, "The Vermont Portfolio Assessment Program: Findings and Implications," *Educational Measurement: Issues and Practices* 13, no. 3 (fall 1994b): 5–16; B. Stecher, "Implementation and Impact of the Vermont Portfolio Assessment Program" (paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, Ga., April 1993). B. Stecher and E. Hamilton, "Portfolio Assessment in Vermont, 1992–93: The Teachers' Perspective on Implementation and Impact" (paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, La., April 1994); B. Stecher and K. Mitchell, "Portfolio-Driven Reform: Vermont Teachers' Understanding of Mathematical Problem Solving" (CSE Technical Report No. 400, Los Angeles, Calif.: CRESST/UCLA, 1995).
7. Kentucky Department of Education, *Kentucky Instructional Results Information System: 1991–92 Technical Report;* Kentucky Department of Education, *Kentucky Instructional Results Information System 1992–93 Technical Report.*

8. Kentucky Department of Education, *Kentucky Instructional Results Information System 1992–93 Technical Report.* R. K. Hambleton, R. M. Jaeger, D. Koretz, R. L. Linn, J. Millman, and S. E. Phillips, *Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991–1994* (Frankfort, Ky.: Kentucky General Assembly, 1995). E. Kifer, "Perceptions, Attitudes, and Beliefs About the Kentucky Education Reform Act (KERA)," in *A Review of Research on the Kentucky Education Reform Act (KERA),* ed. Kentucky Institute for Education Research (Frankfort, Ky.: Kentucky Institute for Education Research, 1994); D. Koretz, S. Barron, K. Mitchell, and B. Stecher, *Perceived Effects of the Kentucky Instructional Results Information System (KIRIS)* (Santa Monica, Calif.: RAND, 1996); R. Pankratz, *Summary of Research Related to KERA* (Frankfort, Ky.: Kentucky Institute for Education Research, 1995); B. Stecher, S. Barron, T. Kaganoff, and J. Goodwin, "The Effects of Standards-Based Assessment on Classroom Practice: Results of the 1996–97 RAND Survey of Kentucky Teachers of Mathematics and Writing" (CSE Technical Report No. 482, Los Angeles, Calif.: CRESST/UCLA, 1998).
9. P. G. LeMahieu, D. H. Gitomer, and J. T. Eresh, "Portfolios Beyond the Classroom: Data Quality and Qualities" (manuscript no. 94-01, Princeton, N.J.: Educational Testing Service, 1995).
10. C. A. Gentile, J. Martin-Rehrmann, and J. H. Kennedy, "Windows into the Classroom: NAEP's 1992 Writing Portfolio Study" (report no. 23-FR-06, Washington, D.C.: U.S. Department of Education, National Center for Education Statistics, 1995).
11. W. H. Thomas, B. A. Storms, K. Sheingold, J. I. Heller, S. T. Paulukonis, A. M. Nunez, and J. Y Wing, "Portfolio Assessment Research and Development Project: Final Report" (Princeton, N.J.: Educational Testing Service, Center for Performance Assessment, December 1993).
12. D. Koretz, "Large-Scale Portfolio Assessments in the U.S.: Evidence Pertaining to the Quality of Measurement," *Assessment in Education* 5, no. 3 (1998).
13. Ibid.
14. Koretz, Stecher, et al. "Can Portfolios Assess Student Performance?"
15. Koretz, Stecher, et al., "Evolution of a Portfolio Program."
16. D. Koretz, "Large-Scale Portfolio Assessments in the U.S."
17. Gentile, Martin-Rehrmann, and Kennedy, "Windows into the Classroom."
18. Hambleton, Jaeger, et al., *Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991–94.*

19. Koretz, Barron, et al., *Perceived Effects.*
20. Koretz, Stecher, et al., "Can Portfolios Assess Student Performance?"
21. Koretz, Barron, et al., *Perceived Effects;* Stecher, Barron, et al., "Effects of Standards-Based Assessment on Classroom Practice."
22. Koretz, Barron, et al., *Perceived Effects;* Stecher, Barron, et al., "Effects of Standards-Based Assessment on Classroom Practice"; LeMahieu, Gitomer, and Eresh, "Portfolios Beyond the Classroom"; Thomas, Storms, et al., "Portfolio Assessment Research and Development Project."
23. Stecher, Barron, et al., "Effects of Standards-Based Assessment on Classroom Practice."
24. Stecher and Mitchell, "Portfolio-Driven Reform."
25. Koretz, Stecher, et al., "The Vermont Portfolio Assessment Program."
26. Koretz, Stecher, et al., "Evolution of a Portfolio Program"; Koretz, Barron, et al., *Perceived Effects*.
27. LeMahieu, Gitomer, and Eresh, "Portfolios Beyond the Classroom."

Chapter 7

Using Performance Assessment for Accountability Purposes

William A. Mehrens

Why is there such great interest in performance assessment? Are large-scale performance assessments administratively feasible, professionally credible, publicly acceptable, legally defensible, and economically affordable?

Performance assessment is currently a hot topic in education, and it is easy to be impressed with the enthusiasm, energy, and optimism displayed by those doing research on performance assessment. However, as with any hot topic, there are those who have put on their advocacy hats before the data support it. It is my hope to bring a reasoned discussion to the issue of performance assessment for accountability purposes.

A simple statement of my position is that I am in favor of performance assessment by individual teachers who integrate their assessments with their instruction; I am in favor of performance

This chapter originally appeared in *Educational Measurement: Issues and Practice* (Spring 1992). It has been slightly revised and updated.

assessment as a supplement to more traditional examinations for licensure decisions;[1] and I am in favor of some *limited, experimental* tryouts of performance assessment for other accountability purposes. Many questions must be answered and problems overcome before it should be used on a wide-scale basis.

The title and thrust of this article are on the use of performance assessment in accountability programs. Yet most of the research and rhetoric regarding the advantages of performance assessment have been in the realm of individual pupil diagnosis. When one switches from local classroom assessment for individual purposes to mandated assessment for accountability purposes, different issues arise. If performance assessment is used for high-stakes accountability purposes, many of the same kinds of problems that have occurred with multiple-choice tests will exist. For example, there will be the potential problem of focusing instruction toward the particular sample of the domain that is being assessed. This will neither be beneficial to instruction nor result in accurate inferences about the domain.

Any assessment used for accountability faces higher criteria than assessment used for individual pupil assistance within the classroom. Any assessment to be used for accountability purposes has to be administratively feasible, professionally credible, publicly acceptable, legally defensible, and economically affordable.[2] In my view, performance assessment will have more trouble meeting several of these criteria than do multiple-choice tests.

As Fitzpatrick and Morrison pointed out in 1971, "There is no absolute distinction between performance tests and other classes of tests."[3] The distinction is the degree to which the criterion situation is simulated. Typically, what users of the term mean is that the assessment will require the examinee to construct an original response. Some people seem to call short-answer questions or fill-in-the-blank questions performance assessments. However, it is more common in performance assessment for the examiner to observe the process of the construction; thus, there is heavy reliance on observation and professional judgment in the evaluation of the response.

The first point that should be stressed is that performance assessment really is not new. It was employed when the Gilead

guards challenged the fugitives from Ephraim who tried to cross the Jordan River:

> "Are you a member of the tribe of Ephraim?" they asked. If the man replied that he was not, then they demanded, "Say Shibboleth." But if he could not pronounce the "sh" and said "Sibboleth" instead of "Shibboleth" he was dragged away and killed. As a result, 42,000 people of Ephraim died there at that time.
>
> (Judges 12: 5–6)

That obviously was a performance examination. I point it out because I heard a speaker at a recent professional meeting say that "performance tests have only been around a couple of years." Even a reading of the 1971 Fitzpatrick and Morrison chapter in the second edition of *Educational Measurement*[4] could have prevented such an inaccurate statement. However, it is true that the popularity of talking about performance assessment as the latest solution to our educational problems is a new phenomenon.

Factors Supporting Performance Assessment

Like all "new" developments, performance assessment is backed by a very large number of people for a variety of reasons. Several of the major reasons are as follows: (a) the old (but largely inaccurate) criticisms of multiple-choice tests; (b) the belief of cognitive psychologists that assessment of procedural knowledge requires formats other than multiple-choice questions; (c) the increased concern that multiple-choice tests delimit the domains we should be assessing; (d) the wide publicity of the Lake Wobegon effect of teaching too closely to multiple-choice tests; and, finally, (e) claims that there are deleterious instructional/learning effects of teaching to multiple-choice test formats. Certainly these five points are related and overlapping, but they will be discussed separately.

Traditional (But Largely Incorrect) Criticisms of Multiple-Choice Tests

There have been three main criticisms of objective paper/pencil tests: They are biased, they measure irrelevant content, and the format demands only the ability to recognize an answer—not to actually work problems. Of course, any of these criticisms can be

true, but they are not necessary concomitants of the multiple-choice format.

Bias

This article is not the place to discuss the bias charge, but much research has been published about that issue. Publishers of high-stakes multiple-choice tests know a great deal about what test construction/analyses steps to take to prevent and/or detect bias. Well-constructed multiple-choice tests generally fare well under psychometrically accepted definitions of test bias.

Irrelevant Content

The issue of content relevance is related in part to the issue of whether the multiple-choice format can only be used for a limited number of educational objectives/goals. But the issues are separable. To give you a flavor of the criticism, consider the following quote: "We're spending hundreds of millions of dollars on tests that don't tell us anything about what kids know or know how to do."[5] While this quote was directed more at existing commercial standardized tests than the objective format per se, the rhetoric stems at least in part from incorrect beliefs about what multiple-choice tests can measure. In addition to the incorrect concern about irrelevant content, there is concern about the lack of total inclusiveness of the content and its lack of perfect match with the curriculum.[6]

There will never be universal agreement about the goals/objectives of education. However, one must keep in mind that standardized multiple-choice achievement test domains are based upon very thorough reviews of existing curricula guides and textbooks. These, one would assume, have been developed and/or adopted because they have some match to the goals of the local schools. Most parents do want their children to learn the content domains sampled by multiple-choice standardized achievement tests.

Measurement of Recognition Only

The criticism that multiple-choice tests measure only recognition is reflected in the following statements:

> Standardized multiple-choice tests have drawn increasing fire as too simplistic, measuring the ability to recognize knowledge rather than the ability to think and solve problems, an important skill in today's jobs.[7]

> It's testing for the TV generation—superficial and passive. We don't ask if students can synthesize information, solve problems or think independently. We measure what they can recognize.[8]

The notion that multiple-choice items cannot measure higher-order thinking skills is unfortunate and incorrect. Over the years, Forsyth has given any number of talks illustrating that multiple-choice achievement test items can tap higher-order thinking skills.[9] If his examples have not convinced the doubtful, they simply are not open-minded—or perhaps they do not think at a high enough level. Look at the sample multiple-choice questions sent to students who register for the SAT. You could not possibly answer those questions without engaging in some problem solving and/or higher-order thinking.

Cognitive Psychologists' Influence

Cognitive psychologists distinguish between declarative and procedural knowledge (or content knowledge and process knowledge). As Snow and Lohman[10] point out, all cognitive tasks require both types of knowledge, but different tasks differ in the relative demands they place on the two. It is generally accepted that some types of procedural knowledge are *not* amenable to multiple-choice types of assessment. The increased (and in my view correct) push for procedural knowledge goals has led to an increase in the attempts to engage in performance assessment. (However, this should not result in a *replacement* of objective tests.)

Over the past decade or so, many individuals have been hypothesizing on "what cognitive psychology seems to offer to improve educational measurement."[11] Snow and Lohman suggest that the implications of cognitive psychology are largely for measurement *research* and that "cognitive psychology has no ready answers for the educational measurement problems of yesterday,

today, or tomorrow."[12] Other researchers generally seem to agree with this assessment.[13] None of these researchers suggest wide adoption of their exploratory research.

Based on his research, Siegler warns us "that even seemingly well-documented cognitive psychological models may be drastically incorrect, and that diagnoses of individuals based on these models could only be equally incorrect . . . the time does not seem ripe to advocate their use in the classrooms."[14]

In spite of the somewhat cautionary tone used above, I am convinced that cognitive psychologists do have something to offer those of us in measurement. However, I, like Snow and Lohman,[15] think that it is primarily in terms of helping measurement specialists to develop new and better theories. We should not jump on any "performance assessment for accountability" bandwagon before those theories are understood much more thoroughly.

Delimited Domain

Partly as a result of cognitive psychologists' influence, there has been increased concern that multiple-choice tests cannot assess all the important domains of educational goals/objectives. Across the decades, measurement specialists have agreed that objective tests cannot adequately cover all objectives. For example, no one believes they are a good way to measure perceptual motor skills. However, as measurement-driven instruction has increased, the concern about the delimitation of the measured domains has increased.

Many important areas can be efficiently assessed via multiple-choice questions. As Weinstein and Meyer make clear in their chapter on the implications of cognitive psychology for testing, many different educational tasks require simple recall, particularly tasks in the lower grades and in introductory courses. Further, experts differ from novices in their knowledge base, and research suggests "that domain knowledge is a necessary but insufficient condition for acquiring strategies and expertise."[16]

Collis and Romberg, advocates of performance assessment in mathematics, admit that multiple-choice items provide "an effi-

cient and economical means of assessing knowledge of and ability in routine calculations, procedures, and algorithms. *All* [italics added] seem to agree that these skills are still an important part of mathematics education."[17]

In spite of my belief in the importance of procedural knowledge and the importance of doing some assessing by other than multiple-choice testing, I remain puzzled by some of the writings regarding this "new" performance testing. Some suggest that multiple-choice tests are indirect and what we need are more direct measures of achievement. But cognitive psychologists focus on processes (such as metacognitions) that are *not* amenable to direct measurement.[18] Some think the problem is that multiple-choice tests do not cover a broad enough domain.[19] But performance tests will access narrower domains—perhaps in more depth.[20] Some are concerned with the curriculum-test mismatch and the efforts of educators to change the curriculum to increase the match—these people generally see measurement-driven instruction as a bad thing. Others are interested in using new assessment procedures to reform the curriculum and they hope there is teaching to the assessment. All of this confusion is compounded by those who refuse to separate the issues of test content and test form (which are related, but not identical, issues).

Lake Wobegon Effects

High-stakes tests can lead to teachers' teaching too closely to the test, thus raising scores without raising the inferred achievement. Some advocates of performance assessment suggest that it is appropriate to teach directly to that type of assessment because the instructors will be teaching appropriate material in ways they ought to be teaching it. Consider the following quotes:

> Teaching to these [California Assessment Program] tests is what we want because the tests are 100 percent connected with real-world on-the-job performance.[21]

> If schools spend three or four weeks a year teaching to a performance-based test, at least they'll be teaching things they ought to be teaching in ways they ought to be teaching it.[22]

However, those who feel that performance assessment is the solution to teaching to the test are sadly mistaken. Their reasoning misses the point about inappropriate test preparation. They basically ignore the domain/sample problem that is exacerbated when one delimits the sample as one must in a performance assessment.

Deleterious Instruction

Tied to all the above issues is the apparent belief by some that if one tests via a multiple-choice test and teaches so that students do well on the multiple-choice test, the instruction must be deleterious, but if one assesses via performance measures, the instruction will be beneficial.

It is true that the format of the assessment will have some effect on instructional practices, that this effect will be greater if the assessment is for high-stakes accountability decisions, that answering multiple-choice questions is not a task that is done a lot outside of school, and that excessive instruction tied too closely to an unrealistic form of assessment is a poor instructional strategy. Nevertheless, it is not true that performance assessment will necessarily lead to high-quality instruction. The Honig and Shavelson ideas quoted above are not necessarily true. The California Assessment Program's five performance items in math[23] are certainly not "100 percent connected with real-world on-the-job performance." Further, teachers could spend time teaching correct answers to these questions without "teaching in ways they ought to be teaching it."

Again, I have perhaps sounded too cautionary. Writing assessment has probably increased the instruction of writing and that is a good thing. I suspect performance assessment of safety procedures in the science laboratories might increase the efforts of teachers to teach safety procedures, and that would be a good thing. But it is important to keep in mind Linn's admonition that we need to do more than just assume that the alternatives to multiple-choice items will have no bad side effects of their own.[24] We must be prudent in our charges regarding the ills of multiple-choice test and in our claims about the wonders of performance assessment for instruction.

Problems with Performance Assessment for Accountability

Like other forms of assessment, the particular problems that are likely to be faced with performance assessment vary somewhat depending on a variety of dimensions, such as (a) secure vs. nonsecure assessments, (b) matrix sampling vs. every pupil assessment, and (c) accountability vs. instruction.

Secure vs. Nonsecure Instruments

One disadvantage of performance assessment is that with only a few questions, there is no way to keep the exact content of the exam secure. Once performance assessments have been used, they cannot be reused to test the same higher-order thinking process. One can memorize an answer to a higher-order question just as well as one can memorize an answer to a basic skills question. Thus, performance assessments have to be new each year—adding to the developmental costs and making across-year comparisons of growth very difficult.

Baker and colleagues took a different approach by suggesting that "only if the tasks and scoring criteria are made public . . . can teachers guide students to meet such standards, and then only if the same tasks are used."[25] Although I grant that this may be done without corrupting the inference for some physical performance tests (for example, diving), performance assessment tasks that have a metacognitive component do not allow for such release and reuse of the tasks.

Matrix Sampling vs. Every Pupil Testing

Different cost issues arise with these two methods. Assessments that would be cost prohibitive for every pupil testing may be reasonable in a matrix sampling approach. However, if individual student scores are not reported, this makes the assessments much less useful to individual teachers, and a lack of student motivation makes the results suspect. Further, some high-stakes tasks, such as those used for licensure and high school graduation requirements, demand every pupil testing.

Accountability vs. Instruction

As mentioned earlier, high-stakes tests used for accountability purposes need to meet what Baratz-Snowden[26] has referred to as the five "apple" criteria: administratively feasible, professionally credible, publicly acceptable, legally defensible, and economically affordable.[27] I maintain that performance assessment is likely to have difficulty meeting many of these standards. Currently, it appears to meet the professionally credible and publicly acceptable criteria, but that could be because it is in the fad stage. More careful scrutiny may change that.

Administratively Feasible/Economically Affordable

Because resources are always limited, the costs of performance assessment must be of great concern. The Educational Testing Service has reported that "one state with a strong commitment to educational assessment found that redesigning its state program around performance tasks would increase by tenfold the cost of the existing state assessment program."[28] Given my belief that most performance exercises are not reusable without distorting the inference, there are some very real questions about the developmental costs in performance assessment for accountability.

Even after performance assessments have been developed, the costs of administering and scoring them are high. Frequently special equipment is needed for administration, and it is not feasible to have enough copies for simultaneous administration. Consider, for example, the four components planned for an assessment of teachers' laboratory skills:[29] a preobservation questionnaire, a preobservation conference, an observation, and a postobservation conference. The observation lasts thirty to forty-five minutes, and observers in the pilot study were trained for three days. All this is certainly expensive. This is not to suggest we should not do performance assessments, but cost-benefit ratios must be considered.

Publicly Acceptable

So far the performance assessment advocates have done a good job with public relations. But, as with multiple-choice tests, once perfor-

mance assessments have been used awhile for accountability purposes and the teachers complain about their lack of validity for accountability inferences, there may be a reduction in public acceptability. Once the public understands that the costs will be substantially higher, one might expect some loss of acceptance of the process.

Legally Defensible

"Legally, performance assessment is considered a test."[30] Whether this is how all courts would decide the issue, prudent individuals developing performance assessments for high-stakes decisions would be wise to act as if this were the case.[31] Psychometric experts for plaintiffs generally attack tests based on whether the *Standards for Educational and Psychological Testing*[32] have been followed. One would expect them to do the same for performance assessments. That performance assessments will meet the various psychometric standards of reliability, validity, and so on, has not been adequately demonstrated. Other legal concerns also need to be considered. For example, if there is any disparate impact on protected groups, how might one deal with the fact that observers (graders) may be aware of the group status of the students? If there is debate about the scoring process, will there be documentation of the performance so rescoring can occur?

Professionally Credible

Professional credibility pertains at least to three overlapping groups: teachers, those involved in teacher education, and psychometricians. Because of effective public relations efforts and face validity, performance assessment probably has more credibility than multiple-choice testing for the first two groups. It is impossible to know if that will continue if performance assessment becomes widely used for accountability. Wide use will result in more scrutiny than such assessments have currently been given, and the whole movement could implode following such scrutiny. Psychometricians will probably place or withhold their stamps of approval based on evidence regarding the psychometric properties of the assessments. This may place them on the

credibility continuum at a point different from those individuals who minimize the importance of psychometric properties.

Validity

Generally, psychometricians believe it is important to validate new approaches to testing before any wide implementation.[33] Performance assessments have face validity—or what Popham[34] says can be more pedantically described as verisimilitude. Face validity helps in the acceptance of an assessment procedure, and some level of face validity is essential for public credibility. But, as psychometricians know, face validity "is not validity in the technical sense; it refers not to what the test *actually* [italics added] measures, but to what it *appears superficially* [italics added] to measure."[35] It does not take the place of real validity and is simply not sufficient. To date, there is little evidence on the validity of performance assessments.

In studying the validity of performance assessments, one should think carefully about whether the right domains are being assessed, whether they are well defined, whether they are well sampled, whether—even if well sampled—one can infer to the domain, and what diagnostically one can infer if the performance is not acceptably high.

A wish to assess the different domains was a major reason for implementing performance assessment, and in a general sense, I am in favor of what cognitive psychologists and reform educators are stressing. Nevertheless, the appropriateness of performance domains is as subject to debate as are those domains assessed via multiple-choice tests. In general, performance assessment measures a narrower domain than multiple-choice testing, but assesses it in more depth. Is this good? There should probably be more discussion regarding just which narrow domains need to be assessed in depth.

If one is satisfied that the right domains are being assessed, one should still consider whether they are defined tightly enough. Critics of standardized tests have suggested that the domains are not well enough defined in those tests. My general observation is that the domains of multiple-choice achievement tests that have

been used for accountability purposes have been more tightly defined than many performance assessment domains.

The major problems for valid performance assessment relate to the limited sampling and the lack of generalizability from the limited sample to any identifiable domain. One of the generally accepted advantages of multiple-choice testing is that one can sample a domain very thoroughly. Because performance assessment takes more time, fewer tasks (questions) can be presented. Thus, the sampling of the domain is less dense. For example, in California there were only five mathematics items on the state performance assessment.[36] One would be hard-pressed to generalize to any curricular domain from such a limited sample.

Even if sampling is adequate, there is the question of whether one can generalize from the sample to a larger domain. This is dependent upon the intercorrelations between the portions of the domain in the sample and those portions not in the sample. Certainly research has indicated that higher-order thinking skills and problem solving are specific to relatively narrow areas of expertise, and there appears to be little transfer from one subject matter to another on these constructs.[37]

But even *within* a subject matter area, generalizability is "iffy." As Herman has pointed out, "Research in performance testing demonstrates how fragile is the generalizability of performance."[38] She gives as one example the research that indicates writing skill does not generalize across genres. The teacher's guide for the *Writing* supplement of the *Iowa Test of Basic Skills* reports correlations between essays in different *modes* of discourse that average .36.[39]

Or consider the generalizability of performance in a science laboratory assessment. Some research has been conducted in California on the development of a science laboratory assessment for new teachers. In their 1990 final report, Wheeler and Page wisely state that they do not know if their prototypic exercises will generalize

> across different science laboratory situations—grades K–12; earth, life, and physical sciences; various types of lab activities; different groups of students; and different lab setting, including field trips. . . .

> Conclusions about the generalizability of the assessment should be based on a large-scale field testing that includes many more types of situations.[40]

At this point, we simply do not have enough data indicating the degree to which we can generalize from most of the performance assessments that are being conducted. Much of the evidence we do have suggests that generalizability is extremely limited.

Even if the domain is the correct one, it is well defined, the sample is adequate, and generalizability is possible, validity problems remain. As mentioned earlier, if the assessment is not secure, students will be taught how to do that particular task. This not only makes the inference to the domain inappropriate, it means one may make an incorrect inference about the sample performance. For anything other than a completely physical skill (for example, diving), one is typically making an inference about the cognitive processes used. But one can memorize reasons as well as facts. Anytime one wishes to infer something like a metacognition, it is important that the assessment be secure.

And finally, a threat to validity that deserves mention is the impossibility of making a precise inference from a poor score on a performance assessment. If, for example, one accepts Anderson's theory of skill development, there are three stages: the declarative stage, the knowledge compilation stage, and the procedural stage.[41] At which stage is an individual whose skill development is inadequate? Multiple-choice tests could assess the first two levels.

Reliability

There are several threats to reliability in performance assessment. One has to do with the small number of independent observations (the sampling problem discussed above). A second has to do with the subjectivity of the scoring process. A third has to do with a lack of internal consistency that influences generalizability (discussed above).

Reliability refers to random error in a measurement, and if random error is too great, any perceived relevance of the assessment is illusory because nothing is being measured.[42] Thus, one cannot possibly make any valid inference from the data.

The only performance assessment area that has reported much evidence on reliability has been writing assessment. There, the major evidence reported is rater reliability. It generally runs in the low .80s, although it can be substantially lower or higher. For example, the average inter-rater reliabilities for the *Writing* supplement to the *Iowa Test of Basic Skills* was .95.[43] Welch obtained inter-rater reliability estimates of .75 to .77.[44] Dunbar and colleagues report on nine different studies where the rater reliabilities range from .33 to .91.[45] To obtain the higher levels of rater reliability is costly. It requires careful selection and extensive training of the raters, precise scoring guidelines, and periodic rechecking of rater performance.

Score reliabilities are reported less often, but, when reported, are quite a bit lower. The *Writing* supplement to the *Iowa Test of Basic Skills* had an average score reliability (two essay samples using the *same mode* of discourse) of .48.[46] The score reliabilities reported in the Dunbar article ranged from .26 to .60. As they stated, these values are "extremely low relative to common standards for high-stakes tests."

Given that writing assessment is the most developed and researched mode of performance assessment, it seems safe to conclude that there are serious problems with the reliability of many performance assessments.

Many issues arise concerning scoring, scaling, equating, and aggregating data:

1. It is obvious that there is subjectivity in assigning the scores to a performance. This means that *who* does the scoring is very important for any test used for accountability. Some telling data regarding scoring by anyone having a vested interest in the results come from the judgments of teacher performance by principals. State after state has obtained very negatively skewed distributions when principals score teacher performance. When assessing for *accountability* purposes, it is imperative to have performances scored by those who do *not* have a vested interest in the outcome. Having teachers score their own students' performances will

not work. Further, if the school building or school district is being held accountable for the scores on performance assessments, the scores must come from outside the district.

The issue of *what* is to be scored is also of considerable importance. Typically, "an examinee response is complex and multifaceted, comprising multiple, interrelated parts."[47] One can use either componential or holistic scoring. As Millman and Greene pointed out, in either case, to develop the scoring criteria requires a clear understanding of what it means to be proficient in the relevant domain (which, in turn, assumes there is a good definition of the domain). Most advocates of performance assessment probably will opt for developing scoring profiles.[48] The *Standards* require that the reliabilities of the subscores are reported. Further, if the data are going to be used for individual diagnostic purposes, one should report the reliability of the difference scores in the profile. These will obviously be lower than the individual score reliabilities. The profiles for students' performances may well be so unreliable that they are useless.

2. Determining how to scale the data from performance assessments is another challenge. In his article on the NAEP Proficiency Scales, Forsyth[49] convincingly argues that those scales do *not* yield valid criterion-referenced interpretation. Large-scale performance assessments will likely be equally difficult to scale.
3. Because performance assessments yield fewer independent pieces of data and because specific assessments should not be reused, the equating problems are formidable. For longitudinal comparisons and fairness in accountability, the scores on different forms of performance assessments must be equated so that they represent the same level of achievement regardless of when the performance is assessed, which tasks are given, or which raters score the performance.
4. Decisions about the unit of reporting will be difficult to make. Certainly for those performance assessments that are

based on group activities, the unit cannot be the individual.[50] However, other types of assessment may lend themselves to individual reporting.

Ethnic Group Differences

One of the reasons for moving to performance assessments is that some individuals are hopeful that performance assessments will show smaller ethnic group differences than do multiple-choice tests. The results are not yet all in with respect to this hope, but evidence on writing assessments across the nation do not show smaller differences between black and white performers than are obtained from multiple-choice tests.[51] Further, the data will be more complicated to interpret due to the subjective scoring processes and the potential opportunity (where performance is observed) for scorers to allow ethnicity to influence their scores.

Conclusions/Implications

As measurement specialists have known for decades, multiple-choice tests measure some things very well and very efficiently. Nevertheless, they do not measure everything, and their use can be overemphasized. Performance assessments have the potential to measure important objectives that cannot be easily measured by multiple-choice tests.

Some exciting research has been conducted regarding performance assessment, but much more research is needed. Like Wolf and colleagues, I would call for "mindfulness"[52] in the performance assessment research and hope that the researchers would "be as tough-minded in designing new options as [they] are in critiquing available testing."[53] Evidence regarding psychometric characteristics must be gathered. One cannot "pursue these new modes of assessment . . . on the mere conviction that they are better."[54] And finally, I agree with Wolf and colleagues that researchers should be "standing on the shoulders rather than the faces of another generation."[55]

While continuing the research, performance advocates should not be overselling what performance assessment can do. Wiggins has suggested, "It's wrong to say [performance assessments] were oversold; they were overbought."[56] I do not see it that way. I think they have been both oversold and overbought.

Most large-scale assessments have added performance assessments to their existing array of efficient multiple-choice tests, not replaced them. There is no question but that the multiple-choice format is the format of choice for many assessments—especially for measuring declarative knowledge.

From at least one point of view, performance assessment is a good thing for measurement specialists and education in general. It has resulted in more money and more resources being devoted to assessment. This has opened up a whole new assessment industry. It should result in more research regarding the effects of testing on teaching and learning. Nevertheless, I agree with Haney and Madaus who suggest that "the search for alternatives [to multiple-choice tests] is somewhat shortsighted."[57] We also need to keep in mind a statement Lennon made more than two decades ago:

> To encourage the innocent to root around in the rubble of discredited modes of study of human behavior, in search of some overlooked assessment "jewels," is to dispatch a new band of Argonauts in quest of a nonexistent Golden Fleece.[58]

Finally, we should heed the wisdom of Boring: "The seats on the train of progress all face backwards; you can see the past but only guess about the future."[59]

Notes

1. This is because of the high costs of false positives in licensure.
2. J. Baratz-Snowden, ed., *RFP-National Board for Professional Teaching Standards* (Washington, D.C.: National Board for Professional Teaching Standards, 1990).
3. R. Fitzpatrick and E. J. Morrison, "Performance and Product Evaluation," in *Educational Measurement,* 2d ed., ed. E. L. Thorndike (Washington, D.C.: American Council on Education, 1971): 238.

4. Ibid., 237–70.
5. Albert Shanker, cited in G. Putka, "New Kid in School: Alternate Exams," *Wall Street Journal* (November 16, 1989): B1.
6. See E. L. Baker, M. Freeman, and S. Clayton, "Cognitive Assessment of History for Large-Scale Testing" in *Testing and Cognition,* ed. M. C. Wittrock and E. L. Baker (Englewood Cliffs, N.J.: Prentice Hall, 1991), 131–53.
7. E. R. Fiske, "But Is the Child Learning: Schools Trying New Tests," *New York Times,* 31 January 1990, B1–B2.
8. Linda Darling-Hammond as quoted in Fiske, "But Is the Child Learning," B8.
9. See, for example, R. A. Forsyth, "Measuring Higher-Order Thinking Skills" (presentation at the meeting of the Institute for School Executives, Iowa City, Iowa, 1990).
10. R. E. Snow and D. F. Lohman, "Implications of Cognitive Psychology for Educational Measurement" in *Educational Measurement,* 3d ed., ed. R. L. Linn (New York: American Council on Education and Macmillan Publishing Company, 1989), 263–331.
11. Ibid., 263.
12. Ibid., 320.
13. See S. Ohlsson, "Trace Analysis and Spatial Reasoning: An Example of Intensive Cognitive Diagnosis and Its Implications for Testing," 251–96; A. Lesgold et al., "Applying Cognitive Task Analysis and Research Methods to Assessment," 325–50; and R. L. Linn, "Diagnostic Testing," 489–98, all in N. Frederiksen et al., *Diagnostic Monitoring of Skill and Knowledge Acquisition* (Hillsdale, N.J.: Lawrence Erlbaum Associates, 1990).
14. R. S. Siegler, "Strategy, Diversity and Cognitive Assessment," in *Educational Researcher* 18, no. 9 (1989): 15–20.
15. Snow and Lohman, "Implications of Cognitive Psychology for Educational Measurement."
16. C. E. Weinstein and D. K. Meyer, "Implications of Cognitive Psychology for Testing: Contributions from Work in Learning Strategies," in *Testing and Cognition,* ed. Wittrock and Baker, 42.
17. K. Collis and T. A. Romberg, "Assessment of Mathematical Performance: An Analysis of Open-Ended Test Items," in *Testing and Cognition,* ed. Wittrock and Baker, 102.
18. Weinstein and Meyer, "Implications of Cognitive Psychology for Testing," in *Testing and Cognition*, ed. Wittrock and Baker.
19. E. L. Baker, M. Freeman, and S. Clayton, "Cognitive Assessment of History for Large-scale Testing," in *Testing and Cognition,* ed. Wittrock and Baker, 131–53.

20. Actually, the evidence regarding whether multiple-choice tests and other assessments cover the same domains is quite mixed. Some research suggests the same domains/constructs are being measured; other research suggests that there are some differences. See T. A. Ackerman and P. L. Smith, "A Comparison of the Information Provided by Essay, Multiple-Choice, and Free-Response Writing Tests," *Applied Psychological Measurement* 12, no. 2 (1988): 117–28; R. E. Bennett, D. A. Rock, and M. W. Wang, "Equivalence of Free-Response and Multiple-Choice Items," *Journal of Educational Measurement* 28, no. 1 (1991): 77–92; M. Birenbaum and K. K. Tatsuoka, "Open-Ended Versus Multiple-Choice Response Formats—It Does Make a Difference for Diagnostic Purposes," *Applied Psychological Measurement* 11, no. 4 (1987): 385–96; R. Farr, R. Pritchard, and B. Smitten, "A Description of What Happens When an Examinee Takes a Multiple-Choice Reading Comprehension Test," *Journal of Educational Measurement* 27, no. 3 (1990): 209–26; M. E. Martinez, "A Comparison of Multiple-Choice and Constructed Figural Response Items" (paper presented at the annual meeting of the American Educational Research Association, Boston, Mass., April 1990); R. E. Traub and C. W. Fisher, "On the Equivalence of Constructed-Response and Multiple-Choice Tests," *Applied Psychological Measurement* 1, no. 3 (1977): 355–70; R. E. Traub and K. MacRury, "Multiple-Choice Versus Free-Response in the Testing of Scholastic Achievement," in *Tests and Trends 8: Jahrbuch der Padagogischen Diagnostik,* ed. K. Ingencamp and R. S. Jager (Weinheim & Basel, Switzerland: Beltz Berlag, 1990): 128–59; W. C. Ward, "A Comparison of Free-Response and Multiple-Choice Forms of Verbal Aptitude Tests," *Applied Psychological Measurement* 6, no. 1 (1982): 1–12; W. C. Ward, N. Frederiksen, and S. B. Carlson, "Construct Validity of Free-Response and Machine-Scorable Forms of a Test," *Journal of Educational Measurement* 17, no. 1 (1980): 11–30.
21. Honig, cited in C. Pipho, "Stateline," *Phi Delta Kappan* 71, no. 4 (1989): 262–63.
22. Richard Shavelson, cited in R. Rothman, "States Turn to Student Performance As New Measure of School Quality," *Education Week* 9, no. 10 (1989): 1, 12–13.
23. California State Department of Education, *A Question of Thinking: A First Look at Students' Performance on Open-Ended Questions in Mathematics* (Sacramento, Calif.: California State Department of Education, 1989).
24. S. Moses, "Assessors Seek Test That Teaches," *APA Monitor* 21, no. 11 (1990): 36–37.
25. Baker et al., "Cognitive Assessment of History for Large-Scale Testing," in *Testing and Cognition,* ed. Wittrock and Baker, 137.
26. Baratz-Snowden, *RFP-National Board for Professional Teaching Standards.*
27. Admittedly, her writing pertained to licensure tests, but I believe the generaliza-

tion of the criteria to accountability assessment is reasonable.

28. Educational Testing Service, *Annual Report* (Princeton, N.J.: Educational Testing Service, 1990), 6.
29. P. Wheeler, "Assessment of Laboratory Skills of Science Teachers via a Multi-Methods Approach" (paper presented in the symposium on Innovative Assessment Prototypes for the California New Teacher Project at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Boston, Mass., April 1990).
30. B. R. Nathan and W. F. Cascio, "Introduction: Technical and Legal Standards" in *Performance Assessment: Methods and Applications* (Baltimore, Md.: The Johns Hopkins University Press, 1986), 1.
31. See *Watson v. Fort Worth Bank and Trust,* 1988, for a discussion of this issue in employment testing.
32. AERA, APA, and NCME, *Standards for Educational and Psychological Testing* (Washington, D.C.: American Psychological Association, 1985).
33. See R. S. Nickerson, "New Directions in Educational Assessment," *Educational Researcher* 18, no. 9 (1989): 3–7.
34. See W. J. Popham, "Face Validity: Siren Song for Teacher-Testers," in *Assessment of Teaching: Purposes, Practices, and Implications for the Profession,* ed. J. V. Mitchell Jr., S. L. Wise, and B. S. Plake (Hillsdale, N.J.: Lawrence Erlbaum Associates, 1990), 1–14.
35. A. Anastasi, *Psychological Testing,* 6th ed. (New York: Macmillan, 1988), 144.
36. California State Department of Education, *A Question of Thinking.*
37. See S. P. Norris, "Can We Test Validly for Critical Thinking?" *Educational Researcher* 18, no. 9 (1989): 15–20, for a discussion of both epistemological and psychological generalizability of critical thinking.
38. J. Herman, "Research in Cognition and Learning: Implications for Achievement Testing Practice," in *Testing and Cognition,* ed. Wittrock and Baker, 157.
39. A. N. Hieronymus et al., *Writing Supplement Teacher's Guide: Iowa Tests of Basic Skills* (Chicago: Riverside Publishing Co., 1987), 28.
40. P. Wheeler and J. Page, *Development of a Science Laboratory Assessment for New Teachers, Grades K–12,* Final Report (Mountain View, Calif.: RMC Research Corporation, 1990), 60–61.
41. J. R. Anderson, *The Architecture of Cognition* (Cambridge, Mass.: Harvard University Press, 1983).
42. Fitzpatrick and Morrison, "Performance and Product Evaluation."
43. Hieronymus et al., *Writing Supplement Teacher's Guide.*
44. C. Welch, "Estimating the Reliability of a Direct Measure of Writing Through

Generalizability Theory" (paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1991).

45. S. B. Dunbar, D. M. Koretz, and H. D. Hoover, "Quality Control in the Development and Use of Performance Assessments," *Applied Measurement in Education* 4, no. 4 (1991): 289–303.
46. Hieronymus et al., *Writing Supplement Teacher's Guide.*
47. J. Millman and J. Greene, "The Specification and Development of Tests of Achievement and Ability" in *Educational Measurement,* R. L. Linn, ed., 344.
48. D. Wolf et al., "To Use Their Minds Well: Investigating New Forms of Student Assessment," in *Review of Research in Education,* ed. G. Grant (Washington, D.C.: American Educational Research Association, 1991), 31–74.
49. R. A. Forsyth, "Do NAEP Scales Yield Valid Criterion-Referenced Interpretations?" *Educational Measurement: Issues and Practice* 10, no. 3 (1991): 3–9.
50. See, for example, the prototype math exercises for the Maryland State Department of Education, in Maryland State Department of Education, *Maryland School Performance Assessment Program: Prototype Mathematics Task* (Maryland State Department of Education, 1990).
51. S. B. Dunbar, "Comparability of Indirect Measures of Writing Skill As Predictors of Writing Performance Across Demographic Groups" (paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April 1987); L. Feinberg, "Multiple Choice and Its Critics," *The College Board Review,* no. 157 (1990); R. L. Linn, E. L. Baker, and S. B. Dunbar, "Complex Performance-Based Assessments: Expectations and Validation Criteria," *Educational Researcher* 20, no. 8 (1991): 15–21.
52. Wolf et al., "To Use Their Minds Well," 33.
53. Ibid., 60.
54. Ibid., 62.
55. Ibid., 36.
56. Cited in R. Rothman, "New Tests Based on Performance Raise Questions," *Education Week* 10, no. 2 (1990): 1, 10, 12.
57. W. Haney and G. Madaus, "Searching for Alternatives to Standardized Tests: Whys, Whats, and Whithers," *Phi Delta Kappan* 70, no. 9 (1989): 683.
58. R. T. Lennon, "A Time for Faith" (presidential address at the annual meeting of the National Council on Measurement in Education, Los Angeles, Calif., April 1981), 3–4.
59. E. G. Boring, *History, Psychology and Science* (New York: Wiley, 1963), 5.