COMPLEX, PERFORMANCE-BASED ASSESSMENT: EXPECTATIONS AND VALIDATION CRITERIA

CSE Technical Report 331

Robert L Linn

Center for Research on Evaluation, Standards
and Student Testing,
University of Colorado at Boulder

Eva L Baker

Center for Research on Evaluation, Standards
and Student Testing, UCLA

and

Stephen B. Dunbar

University of Iowa

We thank Leigh Burstein, Dan Koretz, and Lorrie Shepard for their helpful comments on earlier drafts of this paper. An earlier version of this paper was presented at the 1990 ECS CDE Assessment Conference, Boulder, CO, June.

COMPLEX, PERFORMANCE-BASED ASSESSMENT: EXPECTATIONS AND VALIDATION CRITERIA

Abstract

In recent years there has been an increasing emphasis on assessment results as well as increasing concern about the nature of the most widely used forms of student assessment and uses that are made of the results. These conflicting forces have helped create a burgeoning interest in alternative forms of assessments, particularly complex, performance-based assessments. It is argued that there is a need to rethink the criteria by which the quality of educational assessments are judged and a set of criteria that are sensitive to some of the expectations for performancebased assessments are proposed.

COMPLEX, PERFORMANCE-BASED ASSESSMENT: EXPECTATIONS AND VALIDATION CRITERIA

Assessment policies and practices at the local, state, and national levels are in a period of rapid transformation. The direct assessment of complex performances provide the vision that is guiding many of the current efforts to transform assessment. Examples include a strong push to use more open-ended problems, essays, hands-on science problems, computer simulations of realworld problems, and portfolios of student work. Collectively such measures are frequently referred to as “authentic” assessments (e.g., Archibald & Newman, 1988; Wiggins, 1989) because they involve the performance of tasks that are valued in their own right. In contrast, paper-and-pencil, multiple-choice tests derive their value primarily as indicators or correlates of other valued performances.

Unfortunately, indicators are too often confused with goals, just as norms are too often confused with standards. When this happens, the indicator or norm is apt to lose its value. Worse, the processes that may help fulfill the fundamental goal often become distorted. The greater the gap between the indicator and the goal the greater the likelihood of distortion, particularly when too much emphasis is placed on the indicator. This lack of correspondence between indicator and goal has become an increasingly important concern for traditional tests of achievement as the stakes of the assessments have increased and provides substantial motivation for the recent pleas for “authentic” assessment.

Although the call for authentic assessment seems new to many, it has been standard advice from some measurement specialists for a long time. Lindquist (1951), for example, argued that **“it should always be the fundamental goal of the achievement test constructor to make the elemente of his test series as nearly equivalent to, or as much like, the elements of the criterion series as conseguences of efficiency, comparability, economy, and expediency will permit”** (p. 152, emphasis in the original). With regard to the construction of tests intended to measure higher-order thinking and critical reasoning skills, Lindquist (1951) went on to note that "the most important consideration is that the test questions require the examinee to do the same things, however complex, that he is required to do in the criterion situations (p. 154, emphasis in the original). Clearly, questions of validity focus their attention on long-range objectives, criterion situations, if you will, and the extent to which they are reflected in the tasks presented to learners on a test.

Lindquist's views on the proper goals for the constructor of achievement tests add some historical perspective to the current debate on alternative assessment instruments, for

that debate has largely evolved into discussions of local, state, and national levels about long-range objectives and how best to measure them. In that regard, many writers have criticized the limitations of multiple-choice tests (e.g., Archibald & Newman, 1987;

Frederiksen, 1984; Resnick & Resnick, in press; Shepard, in press). Others point to advances in cognitive and developmental psychology that may offer the possibility of linking assessment to theories of how learning occurs (e.g., Glaser, 1984; Greeno, 1989; Snow, 1980; Snow & Lowman, 1989). Still other writers and policymakers have called for reform and accountability focusing their attention on what our children know, don't know, and should know. One prominent call for educational reform is that

> "By the year 2000, American students will leave grades four, eight, and twelve having demonstrated competency in challenging subject matter including English, mathematics, science, history, and geography; and every school in America will ensure that all students learn to use their minds well, so they will be prepared for responsible citizenship, further learning, and productive employment in our modern economy" (Statement of one of the Nation's educational qoals, Office of Educational Research and Improvement, 1990, p. 23).

With disparate and influential support for alternatives to large-scale assessments as they are presently conceived and implemented, it is not premature to pose questions about the standards of quality alternative assessments ought to satisfy, given that they may become increasingly prominent in assessment programs of the 90s.

Criteria for Evaluating Assessments

Increasingly, alternative approaches to assessment seek to attend to complex learning and processes that are alluded to in this statement of national goals. These assessments appear to have higher fidelity for the goals of instruction than the more familiar indirect measures. Direct assessments of writing, for example, provide instances of the tasks that we would like students to be able to perform whereas questions about proper grammar are designed to measure what are best termed enabling skills as partial indicators of actual ability to write. In short, direct assessments of performance appear to have the potential of enhancing validity.

Of course, appearance, or what in more traditional terms would be called face validity, is not enough. Evidence must support the interpretations and demonstrate the technical adequacy of “authentic” assessments. Skeptics will question the value of new assessments and ask to see the evidence that shows that they are worth the cost, both in terms of dollars and level of effort. But what sort of evidence is needed and by what criteria should these alternatives to current standardized tests be judged?

Relatively few of the proponents of alternative approaches to assessment have addressed the question of criteria for evaluating the measures. Simply because the measures are derived from actual performance or relatively high fidelity simulations of performance it is too often assumed that they are more valid than multiplechoice tests. Many issues concerning the evaluation of the new forms of assessment being developed have not been sufficiently addressed. To develop technically sound performance assessments, portfolios, simulations, etc., we must address certain criteria for evaluating such assessments.

There are, of course, well established psychometric criteria for judging the technical adequacy of measures. Key among these are criteria that stem from the fundamental concepts of reliability and validity, but expanding on their traditional conceptions seems appropriate considering the stated virtues of many new approaches to assessment. Reliability has too often been overemphasized at the expense of validity; validity itself has been viewed too narrowly.

For example, if great weight is attached to the traditional criteria of efficiency, reliability and comparability of assessments from year to year, the more complex and time-consuming performance-based measures will compare unfavorably to traditional standardized tests. Although efficiency, reliability, and comparability are important issues that cannot be ignored in new forms of assessment, they should not the only, or even the primary criteria in judging the quality and usefulness of an assessment. Nonetheless, they are issues that will require attention and careful design of procedures to assure that acceptable levels are achieved for the particular purposes of an assessment.

Experience with writing assessments provides a good starting place in this regard. For example, the use of sophisticated analytical procedures such as the item response theory model for graded scoring of direct writing assessments has proven to be an effective means of constructing scales from judgments of openended responses that have a model-based psychometric foundation (Bock ~ McCabe, 1990). Similarly, the use of an assessment design that systematically rotates prompts in and out of the pool from year to year so that drifts in grading standards can be taken into account has been found to be an effective way of dealing with the problem of year-to-year comparability of assessments.

Additional work on the designs and analytical procedures in order to enhance efficiency, reliability, and comparability is clearly needed. However, it is, if anything, even more critical that we expand the criteria that we use to judge the adequacy of assessments at the same time we expand the forms of assessments that we use.

Modern views of validity already provide the theoretical rationale for expanding the range of criteria. In practice, however, validity is usually viewed too narrowly and given short shrift in technical presentations of evidence supporting an assessment procedure. Content frameworks are described and specifications for the selection of items are provided for standardized achievement tests. Correlations with other tests and sometimes with

teacher assessments of achievement may also be presented. Such information is relevant to judgments of validity, but does not do justice to the concept. An expanded framework of validity concepts can help clarify the kinds of information alternative forms of assessment offer and thereby help establish complementary roles for conventional and alternative approaches.

What are some of the criteria that are suggested by a broader view of validity and why might these alternative criteria be central to the evaluation of the adequacy of new forms of educational assessment? Although the following set of proposed criteria is not exhaustive, it provides a framework that is consistent with both current theoretical understandings of validity and the nature and potential uses of new forms of assessment.

Consequences. Messick (1989) has argued that validation involves the development of a consequential basis for test score interpretation and use as well as the more traditional evidential basis. If an assessment program leads teachers to spend more time on concepts and content included in the assessment and less time teaching content that is not included on the assessment, for example, these are consequences that must be taken into account in judging the validity of the interpretations and uses of the assessment results. Similarly, if an assessment leads to the heavy use of practice materials that closely match the format of the assessment, that is again a consequence that must be taken into consideration.

Although theoreticians such as Messick (see, also Cronbach, 1988) have stressed the criticality of qiving attention to the consequential basis of validity, prior to the recent pleas for authentic assessment, consequences could rarely be listed among the major criteria by which the technical adequacy of an assessment was evaluated. If performance-based assessments are going to have a chance of realizing the potential that the major proponents in the movement hope for, it will be essential that the consequential basis of validity be given much greater prominence among the criteria that are used for judging assessments.

High priority needs to be given to the collection of evidence about the intended and unintended effects of assessments on the ways teachers and students spend their time and think about the goals of education. It cannot just be assumed that a more "authenticn assessment will result in classroom activities that are more conducive to learning. We should not be satisfied, for example, if the introduction of a direct writing assessment led to great amounts of time being devoted to the preparation of brief compositions following a formula that works well in producing highly rated essays in a 20 minute time limit.

A similar statement could be made in the area of mathematics assessment. Consider, for example, the following mathematics question from the California Assessment Program.

"James knows that half of the students from his school are accepted at the public university nearby. Also, half are accepted at the local private college. James thinks that this adds up to 100%, so he will surely be accepted at one or the other institution. Explain why James may be wrong. If possible, use a diagram in your explanation" (Pandy, 1990, p. 50).

Items such as the above are consistent with conceptualizations of mathematics such as those articulated in Curriculum and Evaluation Standards for School Mathematics (National Council of Teachers of Mathematics, 1987) that emphasize problem solving, the notion that there are many ways of solving problems rather than a single right answer or algorithm to be memorized, and the communication of mathematical ideas.

In the discussion that accompanies the above question, Pandy (1990) identifies several desirable question characteristics. Included among these are the notions that they should "serve as exemplars of good teaching practices that are not likely to distort the teaching and learning process" (p. 45). In addition, he argues that such questions should "not be directly teachable, however, teaching for them will result in good instruction" (Pandy, 1990, p. 45). These are highly desirable consequences. But accomplishing these outcomes may depend on a variety of factors in addition to the nature of the problems that are posed for students. If such problems became a standard feature of an assessment used to hold teachers accountable, for example, highly rated solutions of the above problem or of quite similar problems could be provided to students and students could memorize those solutions without developing the type of thinking and problem solving skills that are intended. But, in any case, evidence regarding both the intended and unintended effects on teaching practises and student learning is needed.

Portfolio assessments also raise interesting questions regarding the consequential basis of validity. What constitutes a portfolio can vary widely from one setting to another. Random samples of student work may be assembled in some cases. Perhaps more typical, however, are the portfolios used in one Iowa school district, in which students and teachers make collaborative decisions about each piece to be included. If this latter version of the portfolio represents the student's best work (it may or may not), and if students engage in creative endeavors in part so that they have materials to include in the portfolios, it would seem important to know, for example, how much time during the school year is spent perfecting the entries. The extent to which time is influenced by the way the portfolios are used is also of interest. In this vein, one might reasonably inquire about whether breadth of a student's activities will suffer from overemphasis on a few entries.

Considering validity in terms of consequences forces our attention on aspects of the assessment process that may not be intended or anticipated by the designers of the instruments. We know from experience that results from standardized tests can be corrupted, and have clear examples of some of the factors that lead to that corruption (Shepard, in press). It should not be assumed that new forms of assessment will be immune to such influences.

The concepts of directness and transparency proposed by Frederiksen and Collins (1989) are relevant to the consequences criterion. Directness and transparency are thought to be important characteristics of an assessment because of the presumed effects they have on

teaching and learning. It may be argued, for example, that directness is important because focusing on indirect indicator measures may distort instruction. The case of multiplechoice questions about writing versus direct writing samples illustrates this point. Similarly, transparency is considered important because understanding the basis on which performance will be judged facilitates the improvement of performance. In short, both directness and transparency are presumed to be means to the end of more desirable educational consequences. But evidence is needed that these apparent characteristics of an assessment have those intended effects without at the same time having undesirable unintended effects.

Fairness. The criterion of fairness needs to be applied to any assessment. Judgments about the fairness of an assessment, however, are apt to depend heavily on the uses and interpretations that are made of the assessment results. On a non-threatening assessment such as National Assessment of Educational Progress, for example, it is reasonable to include calculator-active problems even though student access to calculators may be quite inequitable. On the other hand, equitable access would be an important consideration in a calculator active assessment used to hold students or teachers accountable.

It would be a mistake to assume that shifting from fixedresponse standardized tests to performance-based assessments would obviate concerns about biases against racial/ethnic minorities or that such a shift would necessarily lead to equality of performance. Results from the National Assessment of Educational Progress (NAEP), for example, indicate that the difference in average achievement between Black and White students is of essentially the same size in writing (assessed by open-ended essays) as in reading (assessed primarily, albeit not exclusively, by multiple-choice questions). In the 1988 assessments effect sizes computed by subtracting the mean for Black students from the mean for White students and dividing the difference by the White student standard deviation, were .70, .65, and .67 in reading at grades 4, 8, and 12, respectively (means and standard deviations obtained from Langer, Applebee, Mullis, & Foertsch, 1990, p. 115).

The corresponding effect sizes for writing were .72, .68, and .62 at grades 4, 8, and 12, respectively (means and standard deviations obtained from Applebee, Langer, Jenkins, Mullis, & Foertsch, 1990, p.118).

In a similar vein, Feinberg (1990) has noted that the addition of a performance section to the California Bar exam in 1984 did not reduce the difference in passing rates between White and minority test takers. Feinberg went on to report that analyses conducted by Stephen P. Klein indicate that “if the performance test scores were adjusted to take into account the lower reliability of performance test grading, the racial gap would be even wider than on the multiple-choice test” (1990, p. 15).

Gaps in performance among groups exist because of difference in familiarity, exposure, and motivation on the tasks of interest. Substantial changes in instructional strategy and

resource allocation are required to give students adequate preparation for complex, time-consuming, open-ended assessments. Providing training and support for teachers to move in these directions is essential. But validly teaching for success on these assessments is a challenge in itself and pushes the boundaries of what we know about teaching and learning. Because we have no ready technology to assist on the instructional side, performance gaps may persist.

The key point is not dependent on the relative magnitude of group differences, however. Regardless of the relative size of the difference in performance between particular pairs of groups it is clear that questions of fairness will loom as large for performance-based measures as they do for traditional tests.

Questions of fairness arise not only in the selection of performance tasks but in the scoring or responses (Haertel, in press, Sackett, 1987). As Stiggins (1987) has stated, “it is critical that the scoring procedures are designed to assure that performance ratings reflect the examinee's true capabilities and are not a function of the perceptions and biases of the persons evaluating the performance.” (p. 33). The training and calibrating of raters is critical in this regard.

Statistical procedures that are currently used by test publishers to identify items that function differently for groups defined by gender or racial/ethnic group membership that are matched for overall test performance (e.g., Wainer & Holland, in press) have utility for flagging items that may need to be eliminated or, at least, submitted to additional review before they are used. Differential item functioning (DIF) is not synonymous with bias, however (Cole & Moss, 1989; Linn, in press). Moreover, as Dorans and Schmitt (1990) has recently noted, DIF procedures rely on the availability of performance on a sizeable number of items that can be used as the matching criterion in judging each individual item. Therefore DIF procedures are not likely to be directly applicable to performance assessments where the number of separate tasks is quite small. One promising development in this area is a technique devised by Welch and Hoover (1991) that can be used for DIF analysis of polychotomously scored items when an independent matching variable is also available. Assessments that combine performance-based measures and fixed-response items may thus be amenable to DIF analysis; however, it seems likely that greater reliance on judgmental reviews of performance tasks is inevitable.

Procedures for identifying materials that may be offensive to some groups of students or that are the source of irrelevant difficulty for a student need to be employed for performance assessments as well as for traditional test items. It is known, for example, that prior knowledge has a strong influence on the ability of students to comprehend what they read (e.g., Bransford & Johnson, 1973; Pearson & Sprio, 1980). Recent efforts to make reading assessments more consistent with good instruction in reading have involved the use of longer passages that are of the type that children encounter when reading short stories or

expository passages in textbooks necessarily lead to the use of fewer passages (e.g., Valencia, 1988; Wixson & Peters, 1987). Although this trend may enhance the ecological validity of the tasks, it may also make it harder to achieve balance with regard to group differences in prior knowledge about the topics addressed in the passages. Relevant prior knowledge may need to be assessed (Baker, Freeman, & Clayton, 1991).

Miller-Jones (1989) has argued forcefully that "tests of ability and achievement are in fact context-specific assessment environments with culturally specific content" (p. 363). One approach to dealing with diversity of backgrounds and experiences that is used in standardized tests is to sample a range of different topics, sometimes involving different cultural contexts. Such an approach is, if anything, more difficult to use with performance-based assessments because the time-consuming nature of the problems limits their number. An alternative approach suggested by Miller-Jones (1989) would involve the use of "functionally equivalent" tasks that are specific to the culture and instructional context of the individual being assessed. As he goes on acknowledge, however, it is "exceedingly difficult" to establish task equivalence (p. 363). To say the least, the idea of differential tasks tailored to the individuals being assessed that can be used to make fair, functionally equivalent assessments of performance poses a major challenge for those interested in developing assessments.

Transfer and Generalizability. A major concern that was brought to the forefront by Cannell's (1988) report that almost all states and most districts are reporting achievement results that are higher than the national median is the degree to which achievement on standardized tests provides an accurate picture of student achievement or misleads because the scores are in some way inflated. Scores on a specific collection of test items are of little interest in and of themselves. They become potentially interesting only to the extent that they lead to valid generalizations about achievement more broadly defined.

Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; see also, Brennan, 1983; Shavelson, Webb, & Rowley, 1989) provides a natural framework for investigating the degree to performance assessment results can be generalized. At a minimum, information is needed on the magnitude of variability due to raters and to the sampling of tasks. Experience with performance assessments in other contexts such as the military (e.g., Shavelson, Mayberry, & Rowley) or medical licensure testing (e.g., Swanson, Norcini, & Grosso, 1987) suggest that there is likely substantial variability due to task. Similarly, generalizability studies of direct writing assessments that manipulate tasks also indicate that the variance component for the sampling of tasks tends to be greater than that for the sampling of raters (Hieronymus & Hoover, 1986; Breland, Camp, Jones, Morris, & Rock, 1987).

Shavelson, Baxter, and Pine (1990) recently investigated the generalizability of performance across different hands-on performance tasks in science using tasks such as

experiments to determine the absorbency of paper towels and experiments to discover the reactions of sowbugs to light and dark and to wet and dry conditions. Consistent with the results in other contexts, Shavelson, et al., found that performance was highly task dependent. The limited generalizability from task to task is consistent with research in learning and cognition that suggests that emphasizes the situation and context specific nature of thinking (Greeno, 1989). It seems clear, however, that the limited degree of generalizability across tasks needs to be taken into account in the design of an assessment program, generally, either by increasing the number of performance assessments for each student or using a matrix-sampling design where different performance assessment tasks are administered to separate samples of students.

There is of course the view that transfer, at least across topics within a domain, can be addressed by using a task specification (Baker & Herman, 1983) or item shell approach (Hively, Patterson, & Page, 1968). In such approaches the range of tasks, (e.g., types of problems to be solved, experiments to be conducted, poems to be analyzed), are specified in advance and assessment tasks are created to represent systematically critical dimensions. The logical value of this strategy in instructional planning is clear.

The traditional criterion of reliability is subsumed under the transfer and qeneralizability criterion, but this traditional view of reliability needs to be expanded. Consistency from one part of a test to another or from one form to another similar (parallel) form is insufficient. Whether conclusions about educational quality are based on scores on fixed-response tests or ratings of performance on written essays, laboratory experiments, or portfolios of student work, the generalization from the specific assessment tasks to the broader domain of achievement needs to be justified.

In judging results from traditional standardized tests, we should demand evidence regarding the degree to which the skills and knowledge that lead to successful performance on multiple-choice test questions transfers to other tasks. Evidence of both near and far transfer such as the ability to use skills demonstrated on multiple-choice tests to solve real-world problems is needed. Although the need to be concerned about transfer and generalization are most obvious in the case of multiple-choice tests, it is an equally critical consideration in other forms of assessment as well.

Cognitive Complexity. Critics of standardized tests are quick to argue that such instruments place too much emphasis on factual knowledge and the application of procedures to solve wellstructured, decontextualized problems (see, for example, Frederiksen, 1984). Pleas for higher-order thinking skills are plentiful. One of the promises of performance-based assessments is that they will place greater emphasis on problem solving, comprehension, critical thinking, reasoning, and metacognitive processes. These are worthwhile goals, but will require that criteria for judging all forms of assessment include attention to the processes that students are required to exercise.

It should not simply be assumed, for example, that a hands-on scientific task encourages the development of problem solving skills, reasoning ability, or more sophisticated mental models of the scientific phenomenon. Nor should it be assumed that apparently more complex, open-ended mathematics problems will require the use of more complex cognitive processes by students. The report of the National Academy of Education's Committee that reviewed the Alexander-James study group report on the Nation's Report Card (National Academy of Education, 1987) provided the following important caution in that regard:

"It is all too easy to think of higher-order skills as involving only difficult subject matter as, for example, learning calculus. Yet one can memorize the formulas for derivatives just as easily as those for computing areas of various geometric shapes, while remaining equally confused about the overall goals of both activities" (p. 54).

The construction of an open-ended proof of a theorem in geometry can be a cognitively complex task or simply the display of a memorized sequence of responses to a particular problem, depending on the novelty of the task and the prior experience of the learner. Judgments regarding the cognitive complexity of an assessment need to start with an analyses of the task but also need to take into account student familiarity with the problems and the ways in which students attempt to solve them. Analyses of open-ended responses that go beyond judgments of overall quality can be quite informative in this regard. Whatever the nature of an assessment, however, we should include among the criteria an analysis of the cognitive complexity of the tasks and the nature of the responses that they engender.

Content Quality. Included among the explicit criteria for judging any assessment should be the question of the quality of the content. The content needs to be consistent with the best current understanding of the field and at the same time reflective of what are judged to be aspects of quality that will stand the test of time. More importantly, the tasks selected to measure a given content domain should themselves be worthy of the time and efforts or students and raters. These considerations are especially important in view of the limited sampling that is likely to occur with performance-based measures. Regardless of format, misconceptions can be fostered by poor quality assessments as well as poor quality instructional materials. Systematic judgments of the quality of the tasks, not unlike the content reviews of items secured by many commercial test publishers, are needed from subject matter experts. It would also be useful to provide evidence about the ways in which students interpret the content that is presented.

One strategy to assure content quality of newer assessments is to involve subject matter experts not only in a review of tasks but in their design. For example, in addition to involving award winning teachers in the selection of primary source material in American

History for a newly designed assessment (Baker, Freemen, and Clayton, 1991), the scoring criteria for student performance were developed using contrasts between essays composed by active historians and those produced by teachers and students, following models in Chi, Glaser, and Farr (1988). These expert essays by historians revealed characteristics (such as premise-driven writing and use of prior knowledge) which were transformed into scoring criteria. This approach focuses on the quality of content knowledge displayed in student responses.

Content Coverage. The comprehensiveness, what Frederiksen and Collins (1989) refer to as the scope, of content coverage provides another potential criterion of interest. Performance assessment recognizes the importance of process sampling, giving it primacy over traditional content sampling. But breadth of coverage should not be overlooked. House (1989) reports wide divergence in views of historians about what should be taught (and assessed) in history, a result which suggests a broad representation of content specialists may be desirable in determining content coverage.

As Collins, Hawkins, and Frederiksen (1990) have recently noted, if there are gaps in coverage teachers and students are likely to underemphasize those parts of the content domain that are excluded from the assessment. They illustrate this issue by reference to Schoenfeld's (in press) description of a geometry teacher in New York who had been recognized for superior teaching based on the performance of his students on the Regents geometry exam. Unfortunately, the superior performance of his students was achieved by having students memorize the twelve proofs that might appear on the Regents exam. The lack of adequate content coverage in this example, clearly led not only to misleadingly high scores, but more importantly to a distortion of the instruction provided.

There may be a tradeoff between breadth of content coverage and some of the other suggested criteria. Indeed, it may be one of the criteria by which traditional tests appear to have an advantage over more elaborate performance assessments. Nonetheless, it is one of the criteria that clearly needs to be applied to any assessment.

Meaningfulness. Although somewhat like motherhood and apple pie, a criterion of meaningfulness to students is worthy of attention. One of the rationales for more contextualized assessments that they get students to deal with meaningful problems that provide worthwhile educational experiences. Analyses of the tasks can provide some relevant information for this criterion. However, investigations of student and teacher understandings of performance assessments and their reactions to them would provide more systematic information relevant to this criterion. Furthermore, studies of motivation during large-scale assessments, such as NAEP, related to meaningfulness may shed light on reasons for low performance.

Cost and Efficiency. To be practical, especially for largescale assessments, ways must be found to keep the costs at acceptable levels. On of the great appeals of paper-and-pencil,

multiple-choice tests is that they are extremely efficient and compared to other alternatives, quite inexpensive. With more labor-intensive performance assessments, greater attention will need to be given to the development of efficient data collection designs and scoring procedures.

Conclusion

In summary, serious validation of alternative assessments needs to include evidence regarding the intended and unintended consequences, the degree to which performance on specific assessment tasks transfers, and the fairness of the assessments. Evidence is also needed regarding the cognitive complexity of the processes students employ in solving assessment problems and the meaningfulness of the problems for students and teachers. In addition, a basis for judging both the content quality and the comprehensiveness of the content coverage needs to be provided. Finally, the cost of the assessment must be justified.

These eight criteria are not intended to be exhaustive set. A variety of other criteria could be suggested. The key point, however, is that the traditional criteria need to be expanded to make the practice of validation more adequately reflect theoretical concepts of validity. This expansion is needed not just as a theoretical nicety, but because judgments about the relative merits of the many efforts now underway to move assessments “beyond the bubble” will depend on the criteria that are used and the relative weight that is given to them.

Identifying new criteria or giving different weight to old criteria may strike some as stacking the deck in favor of alternative assessments over traditional fixed-response tests. The issue is not which form of assessment may be favored by a particular criterion, however. Rather, it is the appropriateness and importance of the criteria for the purposes to which assessments are put and the interpretations that are made of the results. It has long been recognized that validity is not simply a property of the measure. A measure that is highly valid for one use or inference, may be quite invalid for another. It does not follow, for example, that because a multiple-choice, paper-and-pencil test has relatively good predictive validity in low-stakes assessment contexts, the test will lead to valid inferences about the achievement of important educational goals when that test is used in a high-stakes assessment context or to wise decisions about the educational system. If the purpose is to support the latter type of interpretation and use, the criteria for judging the assessment must correspond to the purpose, regardless of the nature or form of the assessment.

The arguments, pro and con, regarding traditional and alternative forms of assessment need to give primacy to evolving conceptions of validity if, in the long run, they are to contribute to the fundamental purpose of measurement, the improvement of instruction and learning. An important outcome of the alternative assessment movement is that it challenges the education community at large to reconsider just what are valid

interpretations of any kind of assessment information. The additional criteria for evaluating educational assessments discussed in this paper are intended to stimulate thinking about the complementary roles served by traditional and alternative assessments, and ultimately to clarify the ways in which each contributes to true educational reform.

References

Alexander, L. & James, H. T. (1987). The nation's report card: Improving the assessment of student achievement. Washington DC: National Academy of Education.

Applebee, A. N., Langer, J. A., Jenkins, L. B., Mullis, I. V. S., & Foertsch (1990). Learning to write in our nation's schools: Instruction and achievement in 1988 at grades 4, 8' and 12. Princeton, NJ: Educational Testing Service.

Archibald, D. A. & Newman, F. M. (1988). Beyond standardized testing: Assessing authentic academic achievement in secondary schools. Washington, DC: National Association of Secondary School Principals.

Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitively sensitive assessment of subject matter: Understanding the marriage of psychological theory and educational policy in achievement testing. In M. C. Wittrock & E. L. Baker (Eds.), Testing and cognition, (pp. 131-153). New York: Prentice-Hall.

Baker, E. L. & Herman, J. L. (1983). Task structure design: Beyond linkage. Journal of Educational Measurement, 20, 149-164

Bock, R. D. & McCabe, P. (1990). Toward a timely state NAEP. Unpublished manuscript prepared for the National Assessment Governing Board.

Bransford, J. D. & Johnson, M. K. (1973). Consideration of some problems in comprehension. In W. G. Chase (Ed.), Visual information processing (pp. 383-438). New York: Academic Press.

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

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City, IA: American College Testing Program.

Cannell, J. J. (1988). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the National average. Educational Measurement: Issues and Practice, 1(2), 5-9.

Chi, M. T. H., Glaser, R., & Farr, M. (Eds.), The nature of expertise. Hillsdale, NJ: Erlbaum.

Cole, N. S. & Moss, P. A. (1989). In R. L. Linn (Ed.), Educational Measurement, 3rd edition (pp. 201-219). New York, Macmillan.

Collins, A., Hawkins, J., & Frederiksen, J. (1990). TechnQ1~gybased performance assessments. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, April.

Cronbach, L. J. (1988). Five perspectives on validity argument. in H. Wainer & H. I. Braun (Eds.), Test validity, (pp. 3-17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependahility of behavioral measurements. New York: Wiley.

Dorans, N. J. & Schmitt, A. P. (1990). Constructed response and differential item functioning: A pragmatic approach. Paper presented at ETS Conference, Constructed vs. Choice in Cognitive Measurement. Educational Testing Service, Princeton, NJ, November 30.

Feinberg, L. (1990). Multiple-choice and its critics. The College Board Review, No. 157, Fall.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. American Psychologist, 32, 193-202.

Frederiksen, J. R. & Collins, A. (1989). A system approach to educational testing. Educational Researcher, 18 (9), 27-32.

Glaser, R. (1984). Education and thinking: The role of knowledge. American Psychologist, 32, 93-104.

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

Haertel, E. (in press). Performance measurement. In M. C. Alkin (Ed.), EncYcloPedia of Educaitonal Research, Sixth edition.

Hieronymus, A. N. & Hoover, H. D. (1987). Iowa Tests of Basic Skills: Writing supplement teacher's guide. Chicago: Riverside Publishing Company.

Hively, W., Patterson, H. & Page, S. A. (1968). A "universe defined system of arithmetic achievement tests. Journal of Educational Measurement, 5, 275-290.

House, E. R. (1989). Report on content definition process in social studies testina. (CSE Technical Report No. 309). Los Angeles: UCLA Center for Research on Evaluation, Standards, and Student Testing.

Langer, J. A., Applebee, A. N., Mullis, I. V. S., & Foertsch, M. A. (1990). T.earning to read in our nation's schools: Instruction and achievement in 1988 at grades 4, 8, and 12. Princeton, NJ: Educational Testing Service.

Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), Educational Measurement (pp. 119-184). Washington, DC: American Council on Education.

Linn, R. L. (in press). The use of differential item functioning statistics: A discussion of current practice and future implications. In H. Wainer & P. W. Holland (Eds.), Differential item functioning: Theory and practice. Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educationa, Measurement, 3rd edition (pp. 13-104). New York, Macmillan.

Miller-Jones, D. (1989). Culture and testing. American E~ychologist, 44, 360-366.

National Academy of Education. (1987). Commentary by the National Academy of Education on the Nation's Report Card. Cambridge, MA: Author.

National Council of Teachers of Mathematics. (1987). Curriculum and evaluation standards for school mathematics. Reston, VA: Author.

Office of Educational Research and Improvement. (1990). Grant Announcement: National educational Research and Development Center Program, Washington, DC: U.S. Department of Education.

Pandy, T. (1990). Power items and the alignment of curriculum and assessment. In G. Kulm (Ed.), Assessing higher order thinking in mathematics (pp. 39-51). Washington, DC: American Association for the Advancement of Science.

Pearson, P. D. & Spiro, R. J. (1980). Toward a theory of reading comprehension instruction. Topics in Language Disorders, 1, 71-88.

Resnick, L. B. & Resnick, D. L. (in press). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), [uture assessments: Changing views of aptitude, achievement, and instruction. Boston: Kluwer.

Sackett, P. R. (1987). Assessment centers and content validity: Some neglected issues. Personnel Psychology, 40, 13-25.

Schoenfeld, A. H. (in press). On mathematics as sense-making: an informal attack on the unfortunate divorce of formal and informal mathematics. In D. N. Perkins, J. Segal, and J. Voss (Eds.), Informal reasoning in education. Hillsdale, NJ: Erlbaum.

Shavelson, R. J., Baxter, G. P., & Pine J. (1990). What alternative assessments look like in science. Paper presented at Office of Educational Research and Improvement Conference, The Promise and Peril of Alternative Assessment, Washington, DC: October.

Shavelson, R. J., Mayberry, P., & Li, W. (in press). Generalizability of military performance measurements: Marine Corps Infantryman. Military Psychology.

Shavelson, R. J, Webb, N. M., & Rowley, G. (1989). Generalizability theory. American Psychologist, ~, 922-932.

Shepard, L. A. (in press) . Psychometricians' beliefs about learning. Educational Researcher.

Snow, R. E. (1980). Aptitude and achievement. New Directions for Testing and Measurement, 5, 39-59.

Snow, R. E. & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational Measurement (3rd ed.) (pp. 263-331) . New York: Macmillan.

Stiggins, R. J. (1987). Design and development of performance assessments. ~Educational Measurement: Issues and Practice, 6(3), 33-42.

Swanson, D., Norcini, J. & Grosso, L. (1987). Assessment of clinical competence: Written and computer-based simulations. Assessment and Evaluation in Higher Education, 1;l, 220-246.

Valencia, S. (1988). Research for reforming the assessment of readinS' comprehension. Paper presented at the annual meeting of the American Research Association, New Orleans, April.

Wainer, H. & Holland,P. W. (Eds.), Differential item functioning: Theory and practice. Hillsdale, NJ: Erlbaum.

Welch, C. J. & Hoover, H. D. (1991, April). Procedures for extending item bias detection techniques to polychotomously scored items. Paper presented at the annual meeting of the National Council on Measurement in Education,,Chicago.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, 70, no. 9, 7033-713.

Wixson, K.. K. & Peters, C. W. (1987). Comprehension assessment: Implementing an integrative view of reading. Educational Psychologist, 22, 333-356.