DOCUMENT RESUME

ED 349 823 FL 020 642

AUTHOR	Baker, Eva L.
TITLE	Issues in Policy, Assessment, and Equity.
PUB DATE	Aug 92
NOTE	31p.; In: Focus on Evaluation and Measurement. Volumes 1 and 2. Proceedings of the National Research Symposium on Limited English Proficient Student Issues (2nd, Washington, DC, September 4-6, 1991); see FL 020 630.
PUB TYPE	Viewpoints (Opinion/Position Papers, Essays, etc.) (120)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Educational Change; Educational Policy; Elementary Secondary Education; *Equal Education; *Limited English Speaking; *Student Evaluation

ABSTRACT

National educational reform presents an unprecedented opportunity to combine policy options, the best technological knowledge, and American concerns about equity and fairness. There are three principal concerns regarding equity in assessment of Limited-English-proficient (LEP) and other student populations: (1) if students are not assessed because of a lack of instruments, they will fail to benefit from the presumed desirable effects of assessment; (2) if LEP students are assessed in English on subject matters such as mathematics, their performance will be handicapped to varying degrees by their lack of English skills; and (3) all students must be provided the opportunity to learn. This paper seeks to describe and define alternative assessment and characteristics; to review the evidence in support of alternative assessment or performance-based assessment; to consider the validity of alternative assessment when it is applied under various policy options; and to present an example of research and development in alternative assessment being conducted at the Center for Research on Evaluation, Standards, and Student Testing (CRESST). Responses to the paper by Lorraine Valdez-Pierce and Peter M. Byron are appended. (VWL)

Issues in Policy, Assessment, and Equity

Eva L. Baker
University of California, Los Angeles

National educational reform presents an unprecedented opportunity to combine our boldest policy options, the best technical knowledge, and American concerns about equity and fairness. To date, the intent of the National Education Goals supports the policy goal of high quality education and restoration of American competitiveness. The six Goals (see Appendix) also refer to challenge and accomplishment as required of "all" students. Particularly, in Goal One, focused on children's readiness for school, the explicit acknowledgment of the importance of health care, early education, and parental guidance push the boundaries of educational reform far beyond the schoolhouse door.

In the last year, efforts have been mounted on the national scene to convert the National Education Goals into policy. The appointment of the National Council on Education Standards and Testing, specifically commissioned to focus on Goal Three, has deliberated on the following questions:

> Is it desirable and feasible to have National Standards of Education?
>
> Is it desirable and feasible to have a National System of Examinations?
>
> Can these policies be implemented while respecting the traditions and legal constraints of local educational control and authority?
>
> What structure or mechanisms should oversee the development of these Standards and Assessments?

Within a six month period, a panel of 32 individuals -- senators, governors, congressmen, administration representatives, educators, and other public figures considered these questions.

The assumptions underlying a set of national education standards, in part, grow from the observation of the successes of other educational systems, particularly those of our trading partners in the Far East and in Europe. Most of these countries have some form of national curriculum since education is a centralized function. Many of these countries have histories of national examination systems, where individual students received certificates of proficiency or pass-

ports to higher education linked to their specific educational accomplishments. Despite evidence and argument that the infrastructures of these countries support education in a far different manner than in the United States, prestige of the teaching profession, for instance, and that certain cultures support a set of explicit and early decisions about the track a child will take in school and in life, there are a great many other concerns about importing educational models into the particular United States context.

The United States differs in important ways from most of the countries we believe to have exemplary educational systems. First, the U.S. is much more diverse -- in economics, in culture, in first languages spoken -- than any of our competitors. Second, the population is greater -- more children are in school at a grade level or two than the total population of countries we are supposed to emulate.

Third, structural nature of poverty among some groups works against a school-based educational reform strategy used in other nations.

And last, the United States possesses, as almost all social critics, foreign and domestic, have noted, a set of values in tension that define many important attributes of American life. Whether from a historical or literary approach, these values seem to suffuse American life and at once provide the context for much of the conflict played out in successive policy options.

Values

Think of America, or read social commentary from 150 years ago to the present and become confronted with the idea of fairness. Fairness is a proposition subscribed to by all but defined differently. In an educational context, it is interpreted in terms of opportunity as well as in terms of outcomes. Schools must provide equal opportunity for learning; legal precedent has held that certain tests are biased (unfair) if particular racial and ethnic groups fail in disproportionate numbers. Fairness eludes the schools and, as our student bodies become more diverse, the schools must find ways to deal with children from cultures, languages, and expectations that mainstream America barely understands, if at all. Fairness is also a matter of financing and, to this point, a national educational plan must address inequities in educational resources.

Pluralism is another key value in America, borne of our immigrant history. Freedom of expression, tolerance for modes of living that differ, and respect for individuals from all backgrounds are cultural mantras. In the educational arena, the boundaries of pluralism are being pushed by arguments for multicultural curricula -- even

separate curricula for individual groups. Changing the game from how we interact with one another to a differentiated content in the curriculum is certain to present perplexing policy options in the future.

When Americans describe themselves, it is often in terms of individualism, a value related to pluralism but with a very different slant. We value a person's right to define his or her own personal goals and to pursue them. We like idiosyncrasy and celebrate individual achievements. Our educational system reflects this value by revering an individual teacher's right to conduct classroom activities by his or her own lights. We give students many choices -- of topic and of courses -- so they can ideally fashion part of their own educational experience.

We also believe in the idea of self-renewal, that people can start over, in 80s language, "reinvent" themselves. Thus, any course of action can be changed, failures, ideally can be overcome; class membership is not a permanent state.

Competition, the need to win or be best, is at the heart of the American psyche. It is exemplified in our economic system, in our obsession with sports and awards, in school with the emphasis on grades and comparisons, state with state, or child with child. It is, in part, an explanation for how the psychometrics of the twentieth century developed to differentiate performance among people rather than to describe its characteristics.

Finally, Americans also believe in community, in the importance of our neighbors, in helping, and in providing aid and mutual support.

It takes little pondering to discern that these values create clusters of tension as emphasis upon them differs by group, by goals held, and over time. Yet, it is precisely these values, and adherence to them by different participants in educational policy, which frame the conflict about national educational standards and testing.

Standards

The term standards in education has meant to define the level of desired performance. In the current debate, educational standards have come to describe, in part, what content and skills a student was supposed to know. This change in usage is, in part, a public relations move, to convey the notion of "high standards." But it is also relevant to the issue of local control. Instead of discussing national curriculum, a topic sure to draw heat from a variety of sources, the term standards is somewhat sanitized by its ambiguity. However, when content standards are discussed, that is, what is expected of a

student in mathematics or science, it is in fact curriculum goals that are really being discussed.

There is also considerable discussion about the strategy by which standards get enunciated and ratified. All agree that the standards should be consensual. They should be developed by representative groups of scholars and practitioners and reviewed by teachers, embraced by policy makers, and so on. Some believe that it is best to begin the process from the end, by creating examples of performances that students should exhibit and derive the standards from this set of performances. In an earlier era, this strategy was called backward chaining and worked when one had a good idea of what the goal was to be in the first place. In subject matters where consensus may be difficult to find, for instance, in literature or social studies, this strategy seems less sensible.

The model used by policy makers in the recent discussion of national standards has been the standards developed by the National Council of Teachers of Mathematics. These standards define, in fairly global terms, what is expected of students in mathematics. The standards are notable because they emphasize problem solving and applications of mathematical thinking, such as estimation and measurement. Because these standards were developed with contributions and participation from many major players in mathematics education, they are often held up as an example of what should occur in the other subject matter fields identified in Goal Three of the National Education Goals: language arts, geography, history, and science. Underway at the present time are consensus processes in history and science. Additional efforts, focused on developing common objectives for the National Assessment of Educational Progress (NAEP), are in various stages in language arts and geography, as well as art.

In the development of the report Raising Standards for American Education, the National Council on Education Standards and Testing identified not only content standards, described above, but also performance standards -- designed to provide a common language for describing proficiency. Most controversial and the topic of some acrimonious debate was the topic of delivery standards. Simply stated, delivery standards were to describe the desirable characteristics of schools and educational systems. The purposes of such description were to assure that schools provided reasonable opportunities for students and to permit analyses and explanation of student outcomes, appropriately conditioned by their educational experiences.

National Systems of Examinations

The standards issue pales in comparison to the issue of a national examination system. Proponents of such a system argue for it from a variety of platforms. Some see its value in operationalizing standards for accountability purposes. They see the function of examinations in terms of sanctions for poor performance and rewards for achievement. Others believe, again using the accountability line of argument, that common examinations will permit comparisons among children, schools, and states, and drive, via the value of competition, performance upward. For these proponents, the form of the examination makes little difference, although most agree it should reflect curriculum and standards.

For others, the power of a national system of examinations inheres, in part, in changing dramatically the form of tests administered to students. At issue is the effect of multiple choice tests on the quality of education. Although everyone would be quick to acknowledge that any test has a reductionist function, multiple choice tests have come in for a strong share of criticism. They are blamed for the piecemeal way teaching and learning occurs (presumably modelling from the format of the test) and for hours spent away from real instruction and focused on test taking skills.

The alternative proposed is a seemingly new form of assessment -- assessment that depends upon students completing longer term tasks, such as essays or projects, and engaging in multiple steps. Instead of multiple choice responses, the students construct their answers and display their proficiencies either in their own performance, such as giving a speech, or in a product they have made, such as an essay or a videotape. One characteristic of these alternative assessments is that they are supposed to be intrinsically motivating, a kind of 1990s relevance. They also may encourage the integration of knowledge across the disciplines.

Thus, alternative assessments focus on students' performance on tasks that require extended time, complex thinking, and integration of subject matter learning (Baker & Linn, 1990; Shavelson, 1990; Torney-Purta, 1990). For leaders in the research and policy communities, the recognition that measures of educational achievement should reflect the complexity of learning has created enormous opportunity to reform education through providing a focus on curriculum, staff development, and instructional improvement (Ambach, 1991; California Assessment Program, 1991; Baron, 1990; Resnick, 1990).

Examples of alternative assessments might be as common as an essay examination or might include tasks such as the following:

1. Situate an aquarium in the school cafeteria.

2. Make a pinwheel (sailboat, or kite) and explain how it works.

3. Create a work-readiness portfolio with evidence of writing, teamwork, technology use.

4. Design, justify, and estimate costs for recreational facilities for your neighborhood.

It is clear that to judge the quality of such tasks, observers or raters must be trained to use specific scoring rules and to demonstrate their ability to do so with reliability, validity, and without bias.

Alternative assessment is promulgated as having purposes and uses including staff development, curriculum reform, diagnosis and reteaching of students, student certification, accountability, job selection, and college or other post-secondary admissions. This is a tall order.

Knowledge Base for Alternative Assessment

The research base on alternative or performance assessment has been described elsewhere (Baker, 1990) but, in sum, we know relatively little about the extent to which alternative assessment is successful in meeting the range of goals identified for its use. Three major sources of information are assessments in other countries, assessments in the military, and the field of writing assessment.

In brief, the evidence from the international community has only limited relevance in the United States context. First, no other country has the psychometric standards -- of validity, reliability, and fairness -- that are common in the United States. The guidelines, articulated by the Standards for Educational and Psychological Measurement, would not be met in any other country in the world. Part of the explanation for this is the psychometric perspective and expertise in this country. But another reason is the propensity of Americans to litigate on the grounds of fairness when test results are used for purposes with serious outcomes, i.e., high stakes, for an individual or system. Much of the technical quality concern in assessment is generated as either offensive or defensive measures from potential litigants in testing enterprises.

Although essay examinations are widespread internationally, with scoring schemes that range from explicit to imaginary, with

very few exceptions, e.g., the Netherlands and Israel, school based assessment is focused on written performance. A recent national policy experiment in Great Britain promoted the use of hands-on alternative assessments in their school systems. The early results suggested that this process had many administrative and resource problems. Teachers were apparently unable to devote the specific, detailed attention needed to judge students' responses and simultaneously maintain the order and pace of instruction for those not being tested at the moment. As a result, there is a general regrouping and rethinking of the utility of this approach.

In the context of vocational training and testing, there are tests in Germany which require particular performance, occupation by occupation. These assessments are integrally linked with the apprentice and other training programs available to non-university bound youth. Studies of this system may be useful for future U.S. analysis.

A second source of information comes from a review of performance assessment in use in the military. Although job performance testing occurs in the assigned unit, the military, for reasons of cost, has stopped using some of the major performance testing, particularly the Skills Qualification test. Although considerable research has been conducted on performance assessments in the military, they have been generally focused on predicting proficient performance from other measures (see Wigdor and Greene, 1991). What is clear is that, with sufficient resources, large scale administration of performance assessments is possible. The military tasks, by and large, focus on identification and procedural tasks, and rarely deal with the conceptual, problem solving, or integration tasks that are the goals of more general educational programs. What is also clear is that such assessments can be subject to bias or corruption as well. When quotas are desired, performance ratings can be manipulated. This almost endemic effect of accountability testing is certainly not avoided because of the type of test used -- performance, multiple-choice, or otherwise.

Research on writing assessment provides the third sector from which we can draw inferences about performance or alternative assessments. Evidence suggests that raters can be reliably trained to make complex judgments, and that these judgments can adhere to an explicit set of criteria, rather than simply on judgments of good and poor performance. Raters can also be helped through specific procedures during the scoring process to cleave to the explicit criteria and not succumb to fatigue or socially redefined categories of judgment. These points are essential if one believes that the rating scale should have direct implication for the instructional activities.

Standards for Quality Alternative Assessments

One effort by CRESST has been to generate a first set of criteria to use in the evaluation of performance assessments. Part of these criteria are applied by inspection. One reviews the assessment and makes judgments about the extent to which it exemplifies the standards. These criteria include whether the assessment is meaningful to students and teachers; whether the content assessed is of high quality; whether there is adequate content coverage; and whether the assessment calls for complex cognitions on the part of the learning. External criteria include whether the assessment promotes generalization and transfer, its fairness, and its cost and administrative practicality. Most important is the consequences that using such an assessment has on the quality of learning and schooling, a dimension difficult to measure but one that should be kept in mind. Although these criteria come from many sources, including the writings of Messick and others, we believe that research studies can be operationalized to assess them as new performance assessments are designed.

Before alternative assessment should become a national policy, there are several areas of work to be done, work quite apart from technical standards.

Evidence of Impact

While there is almost astrological belief that improved assessments will magnetically pull teaching and learning into planetary alignment, what is the evidence for such expectations? Some argue that because multiple-choice tests negatively influenced teaching and led to adaptation to increase scores, e.g., training in test-wiseness and a molecularized curriculum, they believe that setting high standards for assessment will exert control on, of a more positive sort, the instructional behaviors of teachers. One commonly cited source of evidence for this assertion is performance in writing assessment. A particular example is the reputed impact of the implementation of the California Assessment Program (CAP) writing assessment. Data from San Diego School District suggest that writing performance has dramatically improved on most types of writing assessed by CAP over the last three years (Raines & Behnke, 1991). Yet, as the Raines and Behnke report suggests, considerable efforts in staff development were made in parallel to the advent of the CAP writing assessment. Furthermore, staff development did not have to start cold. In California, there has been a strong and continuing effort by virtually all major post-secondary colleges and universities to support improved instruction in writing through the California Writing Project. The conceptual and, to some extent, procedural analyses requisite for the design of staff development preceded the CAP writ-

ing assessment by at least a decade. How ready are disciplines other than writing to provide staff development with a coherent conceptual framework and valid delivery system?

Clarify What is Meant by Alternative Assessment

Enormous confusion and a lot of sloppiness exist in the use of terms. What are we talking about? Passion and description are intertwined. Authentic assessment is a case in point. The term connotes assessment "better than your kind," more real and deserving attention. In practice, it could be used to denote assessments that are more contextualized and either simulate or use performance derived from everyday, non-school tasks. Another inference for the term is that the assessment stimulates more genuine and representative samples of student work because it has more implicit meaning to them. This interpretation is rich in research opportunities. Alternative assessment means anything but multiple-choice (and problem true-false) but generally connotes extended and multi-step production tasks. Such tasks inevitably require the use of raters, judges, or their electronic proxies to determine the quality of the student's effort. Performance assessment encompasses both the meanings above and may specifically call up tasks that require either hands-on activity for solution or tasks where the student solution processes (in science) or ephemeral acts (speech-giving) must be observed.

Alternative assessment definitions must include the designation of the type of intellectual skill assessed (such as explanation or problem solving) and how they interact with sexy format changes. A portfolio is not a portfolio is not a portfolio. We need to hurry the process through while a generally agreed upon lexicon emerges.

Procedures for Developing Performance Assessment Need to be Clear and Consequences of Alternative Strategies Tested

Procedures for developing alternative assessments vary widely and are built mostly on trust. At the heart of the question of development are two issues: first, what is being assessed; second, how will the assessment be used? To the first point, if the assessment is to serve in any way as a standard to demonstrate competency for individuals or to provide a mark for system performance, the identification of the intellectual processes and content/situation domains must be identified. Assessments do not teach by themselves. How are teachers to know which types of instructional tasks are likely to prepare students for alternative assessments if the underpinnings of these assessments are not described in terms the teacher can understand. Some explication of the intention and class of performance of which the alternative is an example must be described. This stric-

ture assumes that at least some alternative assessment attempts to provide a general fra ework in which to place students' accomplishments. Task specification seems an obvious option (Baker, Niemi, Aschbacher, Ni, & Yamaguchi, 1991).

The second issue, the purpose for the assessment, forces a consideration of the issue of the representativeness of student performance on alternative assessments. Given the extended time periods and resources used in many alternative assessment, we need to feel that our findings are trustworthy and fairly represent student capability. Research (Shavelson, 1990; Linn, 1991; Baker, et al., 1991), and pronouncements (Hoover, 1991), suggest that task sampling is a major validity issue. Specifically, researchers have found only moderate correlations between a given student's performance over a set of different tasks. This phenomenon may be due to lack of coherent specifications of the performance task domain, lack of coherent instructional experience, or the inherent instability of more complex performance. Recent research shows some prospect for controlling topic variability (Baker, 1992; Shavelson, Gao, & Baxter 1992) but until some replicated insight on this phenomenon can be developed, using performance assessments for individual student decisions is a scary prospect.

Format and Criteria: Two Critical Features of Alternative Assessment

Among practitioners there is a disconcerting tendency to overvalue differences in format, e.g., hands-on, portfolio, multi-step performance, and leave the identification of scoring criteria " 'til later." Alternative formats for performance are certainly the salient elements of performance assessment. The push for authenticity, that is, the context-sensitive nature of the assessment task, is supported by legions of research in cognitive psychology although this view shows some sign of revisionist thinking. Nonetheless, it simply does not make sense to generate tasks without knowing how or whether they can be credibly scored.

How should scoring rubrics be generated? The most frequent strategy seems to be assembling groups of teachers to decide on scoring dimensions. Evidence from our own research suggests that teachers are not good identifiers of criteria for certain aspects of student performance. For example, we found that teacher-generated criteria could not be transferred in training to other teachers. It was only after we analyzed performances of experts in contrast to teachers and students that we were able to develop scoring rubrics that teachers could be trained to use reliably and that showed desired relationships among other types of student performance and teachers' judgments. These criteria include the students' use of prior knowl-

edge, principles, newly acquired information, and avoidance of misconceptions and, to date, they seem to work well in explanation tasks for history and science. Although we believe criteria should be generated or selected at the time the assessment task is developed, comparative research could be conducted on the cost, feasibility, and resulting quality assessments developed with different models.

In addition, within particular fields, such as writing or history, there are ideological differences of opinion regarding which set of criteria should be employed and whether, for instance, every new task requires its own specially crafted set of scoring criteria. Obviously, such issues are researchable, and a team of us are conducting studies assessing the robustness and validity of alternative kinds of scoring criteria.

The importance of identifiable and public criteria cannot be underestimated. Many analysts have distinguished between the need for common criteria for accountability purposes and the use of teachers' idiosyncratic criteria for assessment in their own classrooms. However, it is clear that equity concerns must drive us in the direction of having common understandings and standards for performance for both accountability and instructional purposes if performance disparities are to be reduced. Yet, if students in different schools are being held to vastly different types of performance, equity issues will exponentially increase with performance assessment.

Adult Views are not Student Views of Assessment

Much is made of the meaningfulness and challenge of alternative assessments as a means to renew students' interest and commitment to school. Our research suggests that students are not nearly so entranced as we are with challenging tests. There is evidence that students do not attempt tasks that seem long and hard. Our studies of anxiety show significant negative relationships with performance on alternative assessments and relatively high levels of anxiety. If students are not willing to engage in such tasks, then our efforts to estimate their performance will be thwarted. The lack of student interest may be a transitional problem, ameliorated following exposure to appropriate instruction.

Educational Equity

Alternative assessment will generate bad news in the short run. Our research in history and science show students have extremely low levels of understanding. Performance is low across the board--terrible for simple short answer assessment of knowledge, those elements of the curriculum thought to be supported by the use of multiple choice tests. Performance in complex explanation, for instance,

integrating prior-knowledge with principle-driven explanation is lower still. Students don't know how to do what is expected of them in these tasks, and they report that they have not been taught such tasks in school. The dilemma is that we cannot improve the quality of these tasks, nor even understand much about their properties, until we can conduct research on students with more than a modicum of knowledge. We need to do teaching experiments to document the obvious proposition that instruction can impact alternative assessment performance. Teachers are going to need to be taught.

Massive support is needed to make alternative assessment a successful reform. Students don't perform well on alternative assessments because teachers have not taught them to do so. Many assume that teachers know how to teach complex cognitive skills but do not do so because of inhibiting multiple choice tests, unresponsive administrations, and so forth. I believe that people do what they know how to do. And I imagine that many teachers simply don't know how to approach instruction of the sort we are describing. We can explain their lack of expertise variously, but it is more important that we consider how to remedy it. For new forms of assessment to have a chance, enormous levels of staff development support must be available to practicing teachers. Significant aspects of teacher education programs must be seriously revamped. Such ambitions require resources. Many agencies are grappling with this problem. For example, the state of California is contemplating a major change in assessment and is exploring options to secure adequate support for staff development. Clearly, the state cannot simply down-load staff development responsibilities, including the continuing design and scoring of assessments, to local districts. We may have even a bigger problem, because redesigned staff development assumes we know what we want to teach teachers to do -- an unsupported proposition.

Beyond resources for assessment and staff, systematic development, implementing alternative assessment has additional costs. On the mundane level, teachers have told us they need additional teaching assistant time simply to use and to manage students during alternative assessments themselves, let alone change their teaching strategies. Costs for copying and materials will rise and this set of resource problems crops up just as local school districts are scaling back dramatically in the face of economic downturn and voters' reluctance to support additional costs for schools.

Equity issues are critical for alternative assessment. Equity has been at the heart of many advances in assessment and underscores some arguments against traditional testing (National Commission on Testing and Public Policy, 1990; Baker & Stites, 1991). Yet, almost paradoxically, the alternative assessment movement faces almost paralyzing equity challenges. First, there is a critical need to educate all but especially minority communities about new developments

in assessment. This need is made more intensive by community suspicion that the establishment is once more changing the game and creating a new barrier by moving away from a known method of testing. Second, the very scoring of alternative assessments based, as they are, on students' observed performance (as opposed to products), raises equity concerns. Raters' (or teachers') expectations may be affected by race and ethnicity. Safeguards will need to be put in place and potential bias will need to be assessed and accounted for. Third, disadvantaged students may suffer disproportionately from their teachers' lack of experience in teaching complex tasks if for no other reason than these students will not so frequently be exposed to compensatory experiences in the home. One way to assist in reducing the disparities is to assure that students have been exposed to desired material. Although reports of simple exposure or opportunity to learn are pale reflections of whether students have had useful and sensible instruction, they are far better than nothing. In a state such as California, with a set of clear curriculum frameworks, classrooms can be monitored on their adherence to such blueprints (CAP, 1991). In fact, we have suggested using portfolios as an indicator of curriculum exposure rather than only or even as an outcome measure (Baker & Linn, 1990). Most importantly, reports of student performance should be conditioned by data on instructional exposure. Nonetheless, we can expect the gap between disadvantaged and economically secure students to widen dramatically. The only saving grace is that when the gap in their performance eventually narrows, the results should have deeper meaning. Evidence to date suggests that such gaps are present between certain ethnicities.

Educational Equity and a National System of Examinations

The report, Raising Standards for American Education, speaks to the equity concerns associated with any national system of assessment. The report recommends that no single test be used for any subject matter and grade level. It supports the development of local examinations to assess the national standards and specifies that a national quality control mechanism, consisting of a review board made up of experts, educators, and the public, oversee the quality of the measures.

This oversight is especially critical when any national examination is to be used for accountability purposes, for instance, to assess the quality of particular programs. A major precept, included in the Appendix to the report, specifies that states or clusters of states who wish their assessment reviewed must provide evidence of validity of the assessment for its purpose and equity interests. Specifically, the report says,

The entity (quality control board) will design, in consultation with state and local educators, guidelines for the collection of evidence on system and school delivery indicators, with specific attention to equity protection. Decisions will be made related to the differential need for delivery indicators for different assessment purposes. States will provide such evidence as it becomes available. When evidence of both delivery indicators and validity standards is adequate, the entity will support the use of high-stakes assessment with secondary school students. It is anticipated that the entity will conduct audit studies, by visiting samples of schools, to verify the delivery and equity evidence provided by states.

States will (also) come forward with their plans for assuring equity in assessment design, administration, and use for gender, for special populations, disadvantaged students, and Limited English Proficient (LEP) students for review by this entity.

There are three principal concerns regarding equity in assessment of LEP and other student populations:

- If students are not assessed because of the lack of instruments, they will fail to benefit from the presumed desirable effects of assessment (improved instruction, accountability, and targeting of resources).

- If LEP students are assessed in English on subject matters such as mathematics, their performance will be handicapped to varying degrees by their English skills. The problem is not easily resolved even by assessment through the native language because of the heterogeneity of students and instructional programs for LEP students. Special procedures will need to be developed to take language and culture into consideration for appropriate assessment.

- All students must be provided opportunity to learn.

Conclusion

Because new forms of testing have a fragile research base, come at high cost, and present significant challenges to the educational community, we are going to have to use them wisely. Rhapsodizing on the wonders of these assessments makes no sense without thinking in parallel about real problems: about issues such as what and how information follows the student from grade to grade, school to school, or district to district; about how to get information on student content expertise, intellectual skill, motivation, and group cooperation all from the same assessment; about how technology can rapidly

be employed to make sense of this process; about how we'll know we've been successful. Although many see alternative assessment predominantly in a personal, interactive, and dynamic classroom environment (Wolf, 1990), one challenge to smarter assessment is whether and how to project alternative assessment simultaneously onto the canvas of large scale assessment. Our interest is to design assessments to serve both instructional and accountability needs. We are unlikely to be successful completely but, for certain definitions of accountability, we probably can make progress (see Burstein, 1991) and justify the expenditure in this area. We have begun to design a theory of assessment that permits simultaneous information for both broad policy and teaching uses of assessment (Baker, Freeman, & Clayton, 1991). This parallel attention to policy and teaching purposes radically revises the common litany of assessment--that separate and different measures are always for different purposes.

Appendix

National Education Goals: By the Year 2000:

Goal 1:

Readiness for School: All children in America will start school ready to learn.

Goal 2:

High School Completion: High school graduation rate will increase to at least 90 percent.

Goal 3:

Student Achievement and Citizenship: 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 may be prepared for responsible citizenship, further learning, and productive employment in our modern economy.

Goal 4:

Science and Mathematics: U.S. students will be first in the world in science and mathematics achievement.

Appendix (Continued)

Goal 5:

Adult Literacy and Lifelong Learning: Every adult American will be literate and will possess the knowledge and skills necessary to compete in a global economy and exercise the rights and responsibilities of citizenship.

Goal 6:

Safe, Disciplined, and Drug-Free Schools: Every school in America will be free of drugs and violence and will offer a disciplined environment conducive to learning.

References

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

Ambach, G. (1991, March). Improving curriculum and instruction. Paper presented at the meeting "Educational Assessment for the 21st Century: The National Agenda," of the Center for Research on Evaluation, Standards, and Student Testing, Los Angeles.

Baker, E.L. (1992). The role of domain specifications in improving the technical quality of performance assessment (CRESST deliverable). Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing.

Baker, E.L., Aschbacher, P., Niemi, D., Chang, S.C., Weinstock, M., & Herl, H. (1991). Validating measures of deep understanding of history. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Baker, E.L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), Testing and cognition. Englewood Cliffs, NJ: Prentice-Hall.

Baker, E.L., Niemi, D., Aschbacher, P., Ni, Y., & Yamaguchi, E. (1991). Using cognitively sensitive assessments of history. In E. L. Baker (Ed.), Designing and scoring content assessments in American history. Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E.L., & Stites. R. (1991). Trends in testing in the USA. In Politics of education association yearbook 90. London: Taylor & Francis.

Baker, E.L. (1990). Assessment and public policy: Does validity matter? Paper presented at the American Evaluation Association Annual Meeting, Washington, D.C.

Baker, E.L., & Linn, R.L. (1990). Advancing educational quality through learning-based assessment, evaluation, and testing (institutional grant proposal for OERI center on Assessment, Evaluation, and Testing). Los Angeles: UCLA Center for Research on Evaluation, Standards, and Student Testing.

Baron, J.B. (1990). How science is tested and taught in elementary school science classrooms. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Burstein, L. (1991). Performance assessment for accountability purposes: Taking the plunge and assessing the consequences. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

California Assessment Program (1991). New Integrated Assessment System for California Schools. Los Angeles.

Hoover, H.D. (1991). Some cautions regarding the use of "alternative" assessments in high stakes situations. Presented at the UCLA/CRESST conference Educational Assessments for the Twenty-First Century: The National Agenda, Los Angeles.

Linn, R.L. (1991). Alternative forms of assessment: Implications for measurement. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

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

The National Council on Education Standards and Testing (1992). Raising standards for American education. Washington: U.S. Government Printing Office.

Raines, R., & Behnke, G. (1991). California assessment program direct writing assessment statewide testing results by district and by school 1989-90. San Diego, CA: San Diego City Schools.

Resnick, L. (1990). Assessment and educational standards. Presentation to The Promise and Peril of alternative Assessment Conference, Washington.

Shavelson, R.J., Gao, X., & Baxter, G.P. (1992). Content validity of performance assessments: Centrality of domain specifications. Unpublished manuscript, University of California, Santa Barbara, CA.

Shavelson, R. (1990). Alternative technologies for assessing achievement. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Torney-Purta, J.V. (1990). Measurement of performance in social studies. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Wigdor, A.K., & Green, B.F., Jr. (1991). Performance assessment in the work place, Volume I. Washington: National Academy Press.

Wolf, D.P. (1990). Assessment as an episode of learning: Making new assessments broad enough. Presentation to the Department of Education conference on alternative assessment.

Response to Eva Baker's Presentation

Lorraine Valdez Pierce
Center for Applied Linguistics

Let me begin by saying that I would like to commend Dr. Baker and her colleagues at CRESST for the impressive work being conducted at their Center. I am particularly impressed by the thoroughness of the studies on alternative assessment, even though they are still limited in number.

In her paper, Dr. Baker sets out four tasks for herself. These were:

(1) To describe and define alternative assessment and its characteristics and comment on these;

(2) To review the evidence in support of alternative assessment or performance-based assessment;

(3) To consider the validity of alternative assessment when it is applied under "various policy options;" and

(4) To present an example of research and development in alternative assessment being conducted at CRESST.

Dr. Baker accomplishes these tasks admirably and in her comments raises many problematic issues, among them, the need to define and clarify terms and purposes for alternative assessment and to ensure that alternative assessments are valid for the purposes for which they are used.

I found very little to disagree with in Dr. Baker's paper in regard to general definitions and uses of alternative assessment and problematic issues in ensuring validity. However, as a discussant at this symposium, I welcome the opportunity to expand upon some of the issues raised in the paper and suggest some things CRESST might consider examining in future studies. Because of the time limitations, I will not address what I see as less significant points I might tend to disagree with but, instead, focus on the key issues raised in the paper.

First, I will focus on purposes of alternative assessment, or how it is used. Second, I will address implications of alternative assessment and high-stakes testing for English language learners. Third, I will address the appropriateness and feasibility of using alternative assessment measures in high-stakes testing programs. For this, I

will draw upon our experiences at the Evaluation Assistance Center-East at Georgetown University in assisting local and state education agencies to conceptualize, design, administer, score, and interpret alternative assessment instruments for students acquiring English, including the design of portfolio assessment systems. And fourth, I will propose recommendations for making future studies on alternative assessment more relevant to the linguistic, cultural, and academic needs of students learning English as their second language.

Purposes of Alternative Assessment

Dr. Baker proposes at least six different purposes of alternative assessment and states that these purposes differ in relation to the broader policy context and to the technical demands they place on the quality of the assessment. The purposes she names include what we can refer to as low-stakes uses, such as school reform, instructional improvement, and grading, and high-stakes uses, such as certification (national testing) and selection (college admission). In a key piece, she states that by wanting to combine both low and high-stakes purposes, practitioners tend to confound problematic issues in validity because they confuse the purposes of alternative assessment with those of traditional, student achievement testing.

While I tend to lean in favor of this argument, I also sense some dissension in the field with regard to the purposes or uses to which alternative assessment can be put. Controversy tends to be inevitable when educational innovations are under consideration. On the one hand, in response to increasing dissatisfaction with the limited information provided by multiple-choice achievement test formats, especially with regard to students not yet proficient in English, practitioners are turning to alternative assessment as a tool not only for identification of students of limited English proficiency but also for monitoring student progress on a continuous and frequent basis. One of the advantages which practitioners perceive alternative assessment to have over once-a-year standardized achievement testing is its potential for providing multiple sources of information over time on student progress in language proficiency and content area knowledge.

On the other hand, we know of states that have in the past or are presently attempting to incorporate alternative assessment measures, such as writing samples, into their statewide testing programs. Some of the states that have been using alternative assessment are in our immediate area, such as Virginia, (in the form of the Literacy Passport Test) and Maryland (in its Functional Literacy Test). Both of these states include student writing samples as part of their statewide testing program. The state of Michigan has also determined alternative assessment to be not only possible but feasible for large-scale, high-stakes purposes. New York has demonstrated

that it is feasible to administer performance tests to every pupil in science. This year Connecticut has implemented the first statewide portfolio assessment system in the nation. By applying alternative assessment in their high-stakes testing programs, these states and others are telling the rest of the nation that they are willing to use alternative assessment and performance assessment to determine whether or not students have achieved the skills that the states most want them to learn. These skills include synthesis and application of individual bits of knowledge. States may be using alternative assessment because they have found that whereas students may do well on multiple-choice tests, this is insufficient evidence for concluding that they can also integrate these facts and skills into desired performance outcomes.

What I suggest we need to look at is how states and local school districts are using the data resulting from the use of alternative assessment measures in large-scale testing programs. I think we can safely assume that they are using the results for the same purposes for which they have used traditional, standardized achievement test results. The following five questions come to mind:

(1) Are states using the data to compare students in order to determine program effectiveness? If they are, students may not have received equal opportunities to learn, may not have participated in the same programs, or may be limited in their English proficiency.

(2) Are states and/or school districts using the results to meet grade promotion and graduation requirements? The research indicates that grade retention is not an effective educational practice, especially with minority students.

(3) Are states using the results to track students? Equity issues indicate that tracking is unacceptable and illegal if ethnic/racial tracking results from these practices.

(4) Are states using the results of alternative assessment measures to provide special instructional services to students who did not attain the minimum score? and

(5) Are states using alternative assessment procedures providing guidelines for the participation or non-participation of LEP students in these statewide testing programs?

Depending on the answers to these questions, the validity of the alternative assessment results and the purposes to which these measures are put become terribly important. Yes, alternative assessment will continue to be used in large-scale programs and perhaps in a national assessment system, although, as Dr. Baker has suggested, this

may lead to a headlong rush to use alternative assessment measures which may not be valid for the purpose for which they were designed.

Implications for Students Learning English

When we consider the implications of high-stakes testing using alternative assessment measures for the general student population, test reliability and validity become essential. But when we consider the implications of using these same alternative assessment instruments with students who are not yet fully proficient in English, the reliability and validity of alternative assessment measures become critical. Although no reference is made in her paper to language minority students or to limited English proficient students in particular, I think some of the points made by Dr. Baker can be expanded upon in order to more clearly see the implications for these students.

First, the review of the literature conducted by Dr. Baker revealed data on the generalizability of performance across tasks. These data indicated that variations in task performance seem to be attributable, in addition to the degree to which tasks were comparable, to differences in specific prior knowledge, including the type of instruction received by students. The State-NAEP data seem to indicate that lower student performance in performance assessments may be a result of a lack of appropriate instructional experience. Students differed in the rate at which they attempted the more open-ended types of items. The implication is that students in "disadvantaged classrooms" who were not exposed to instructional experiences demanding complex performance were not as prepared to take the tests as those who were.

In the extensive studies conducted at CRESST on creating valid alternative assessment measures in the content areas, it was also determined that students brought a relatively low level of prior knowledge to the tasks and so performed poorly. In addition, it was noted that the researchers were "concerned about the heavy verbal load these tasks place on students."

I believe the implications for language minority students not yet proficient in English are clear: In addition to a possible lack of prior knowledge in the form of educational experiences and opportunities, the limited English proficient student also brings a lack of English language skills, including knowledge of the culture in many cases. Many of these students are placed in classrooms where complex performance is not expected and alternative assessment techniques are not used, taught, or practiced. Lacking this exposure, students in the process of acquiring English, who are required to take high-stakes tests that employ alternative assessment measures, are put at an additional disadvantage.

When we consider that students not yet proficient in English may be retained in grade or denied a high school diploma as a result of their performance on high-stakes alternative assessment measures, it becomes of paramount importance to either ensure that these students obtain access to the same kinds of instructional experiences that fluent English-speaking grademates have or that they obtain exemptions, waivers, or other considerations due to their language status, such as alternative assessment in the native language.

Feasibility of Alternative Assessment in Large-Scale Testing Programs

I share Dr. Baker's concern for valid alternative assessment and performance-based assessments and agree that such measures require time, conceptual models, and empirical studies to support their validity. However, I do not believe that performance measures are entirely inappropriate for large-scale assessments, given the following conditions:

(1) The purpose of the assessment is clear and the instrument has construct validity;

(2) Steps are taken to reduce cultural bias so that students from linguistically and culturally diverse backgrounds are not unnecessarily penalized;

(3) Procedures are specified for designing, administering, scoring, and interpreting each measure;

(4) Raters are trained in scoring procedures and inter-rater reliability is consistently high; and

(5) Results obtained on alternative assessment measures and traditional standardized achievement tests are used in combination as opposed to using a score from only one type of test or the other.

At the Georgetown University Evaluation Assistance Center-East, we have received increasing numbers of requests for technical assistance on alternative assessment and portfolio design. We have presented workshops on these topics to teachers and administrators in states all over the Eastern half of the United States, including Puerto Rico and the Virgin Islands, as well as at regional and national conferences. We believe that once practitioners are trained in how to determine the focus of the alternative assessment and in how to score and interpret the data, the potential for increasing the validity of the alternative assessment measure increases. We have suggested that portfolio planning committees composed of teachers and other staff clearly define the purposes of their assessment, select alternative assessment measures which they believe will match their

purpose, and identify specific procedures and criteria for scoring and interpreting these measures. These committees also need to consider assigning weights to the relative value of each measure in a student portfolio. When we consider that teachers and administrators are designing the instruments and setting the standards, we can see that alternative assessment lends itself remarkably well to the setting of local accountability standards.

There is little reason to believe that, given the above-mentioned conditions, alternative assessment could not be implemented in large-scale assessment programs. But these are formidable conditions, similar to those specified by Dr. Baker in her discussion on methods for addressing the comparability of alternative assessments. As Dr. Baker notes, research on alternative assessment is still in its infancy, and we have a long way to go before its applicability to high-stakes testing is clear. In light of this, I would like to make some recommendations for ensuring that future research on alternative assessment addresses the needs of language minority students learning English.

Recommendations

Future studies on alternative assessment need to describe not only the purposes of alternative assessment measures and steps taken to ensure their validity but also key characteristics of the students involved in these studies, because not all language minority students are the same. Specifically, studies need to look at:

(1) Language minority students who are not limited in their English proficiency, representing all grades and skill levels;

(2) Students who are learning English across all grades and of varying levels of English language proficiency;

(3) Variations in students' prior educational background and literacy skills;

(4) The types of instructional programs in which students have participated;

(5) The effects of practicing alternative assessment techniques with language minority students having different levels of English language proficiency, from varying educational backgrounds, and who have participated in mainstream and special instructional programs;

(6) The purposes for which the alternative assessment is being conducted, whether for identification, entry into language support programs such as ESL or bilingual education, monitoring student progress, or exiting from language support programs into mainstream classrooms;

(7) The academic language skills needed for success in English language content-area classrooms, such as math and science, and developing alternative measures to assess the development of these skills for students for whom English is a second or additional language;

(8) The collaborative frameworks, such as school-wide portfolio assessment teams, which facilitate exchange of information between ESL, bilingual education, and mainstream teachers on portfolio assessment; and

(9) Innovative, informative staff development programs which enable teachers and school staff to use alternative assessment frequently and well. By taking into consideration student and instructional characteristics and the purposes of the assessment, we can better determine the potential of each alternative assessment measure to meet its purpose.

Response to Eva Baker's Presentation

Peter M. Byron
New York State Education Department

I'm sure all of you would agree that the previous speakers make my task very difficult. I have two very excellent acts to follow and only hope that my fifteen minute commentary will provide a useful addition to what you have previously heard.

It is an honor to comment on Dr. Eva Baker's paper. Regretfully, you have not had an opportunity to read her excellent paper. Take our word that this paper is well worth your time!

However, before sharing my thoughts on the paper, I would like to congratulate the Office of Bilingual Education and Minority Languages Affairs (OBEMLA) at the Department of Education for convening this symposium. As you are aware, alternative assessment is a topic currently under debate in the measurement community. OBEMLA provided us the service of directing our focus to a current measurement concern. The topic is extremely important because this measurement issue has not been limited to the education community. The issues surrounding assessment have become a part of our evening newscasts. Last night's news provided an endorsement of alternative assessment by Tom Brokaw and a photo opportunity for the President as he visited elementary and secondary schools in Maine. National education policy and assessment has moved to the front burner. The Office of Bilingual Education and Language Minority Affairs provided the field with an opportunity to become engaged in the national debate on education policy. Whether symposium participants agree with the testing instruments proposed or with the nature of the outcomes to be measured, each participant has been given a unique opportunity to listen and to engage in discussion of the national policy and return home better prepared to take an active part in the debate which will certainly follow at the local level.

The third service which OBEMLA provided is that the office combined a measurement community issue and a national policy debate and focused these on limited English proficient children. It is important that limited English proficient children be a part of the discussion from the very beginning and not introduced as an afterthought when policy decisions have been made. The assessment implications for limited English proficient children must be discussed at the very beginning and not when assessment practices are in place.

The Role of Identification, Assessment, and Evaluation

Those with a traditional background in bilingual education recognize that most decisions and most programs depend on three very important issues: identification, assessment, and evaluation.

Dr. Baker and Dr. Valdez Pierce spoke of high-stakes testing. All testing with limited English proficient students is high-stakes. Educators of limited English proficient students must realize that whenever testing is discussed, they must listen. Court decisions, state programs and federal programs all depend on testing and ultimately identification, assessment, and evaluation.

When we fail to appropriately identify limited English proficient students, we lose them. The debate about the number of limited English proficient students in this country is not purely academic. Programs are designed on needs and programs will not be designed if the needs are not properly identified. Identification is essential to program planning.

However, once limited English proficient students are identified they must be placed in appropriate programs. Without valid and reliable assessment or placement practices, students are placed in programs which are not designed for their needs. Whereas identification and assessment are important, program evaluation is essential. Without program evaluation standards, we cannot begin to tell the story about how successful our programs are, and we certainly can't modify programs, if modification is needed.

The discussion we have today and our continuing discussion over the next two or three days is extremely important if not critical for us because future services for limited English proficient students will depend on our deliberations. I ask that each of you keep the words identification, assessment, and evaluation in mind as presenters discuss portfolio assessment, assessment in science and assessment in mathematics and question whether these procedures will result in fair, equitable, and appropriate treatment for limited English proficient students. We must focus our attention on the students. Remember, without appropriate identification there are no students. Without appropriate assessment, the students are placed in the wrong programs and without appropriate evaluation, we can't tell the story of what we do or what the students accomplish. Let's examine why alternative assessment is important by reviewing Dr. Baker's paper.

Alternative Assessment

What is alternative assessment? The literature describes alternative assessment as an alternative to standardized testing which arose because opponents thought that among other problems, standardized tests: (1) provided false student information, (2) were biased against certain students, (3) focused on lower level skills, and (4) allowed teachers to teach to the test.

In commenting on Dr. Baker's paper, I will focus first on her style and then on the content of her paper and conclude with the message which I received from her work.

Style

Dr. Baker is at once crisp, concise, frugal and economical in her use of the English language. Her points are made without redundancy and with a certain imagery that is lacking in many research articles of this genre. I sensed from the paper that the author enjoys playing with words and invite each of you to take the time to test my hypothesis when the paper becomes available!

Content

How is the paper written? The author intended to explain the attributes of alternative assessment and provide examples of each. She did an excellent job. Dr. Baker is fair because she provided the proponent's position yet she is inquisitive in that she does not unquestioningly accept the proponent's view. An interesting aside is that Dr. Baker believes that many of the problems of standardized testing are also problems of alternative testing. Even though alternative assessment proponents talk about measuring higher order thinking skills, Dr. Baker notes that they may be focusing on simple order skills and many of the proponents also teach to the test, albeit a different type of test. She provides arguments from research which would call into question some assumptions made by the proponents.

Most research reviews conclude after criticisms, however, Dr. Baker provides a practitioner's perspective on alternative assessment from her role as a test developer who has operationalized what to others is only theory. The paper provides a primer on alternative assessment which would be difficult to match. The sole limitation which was mentioned by Dr. Valdez Pierce is the limited specific focus on limited English proficient students.

Message

What is the message? Although each of us brings a different perspective and receives a different message, I share these thoughts for your consideration when you read the paper.

This first message is that alternative assessment has a lot to offer. Alternative assessment has forced measurement specialists to rethink traditional practices and in this has contributed greatly to the field. However, hoping for a test and closing one's eyes and crossing one's fingers, does not make a test appear. Even though alternative assessment has a lot to offer, there is a long way to go before alternative assessment can be a reality.

The second message is that tests must successfully undergo a rigorous review by technical standards before they are deemed acceptable. The development of a test is not complete until it undergoes this review. Alternative assessment instruments must be held to technical standards. It may be that alternative assessment instruments and procedures may be more appropriate in areas other than high-stakes testing of limited English proficient students where identification and placement decisions which impact on a student's life choices are made. Alternative assessment may be more at home as a tool for classroom assessment.

In conclusion, because it appears my time is almost exhausted, I would like to shift focus and talk about future considerations. It would be shortsighted if this discussion were finished at the end of this symposium. For that reason, I propose:

1. A task force be convened by the Office of Bilingual Education and Minority Languages Affairs to discuss the testing practices and procedures used to identify and place limited English proficient students and evaluate educational programs serving them. This task force would be particularly charged to ensure that testing and evaluation standards be implemented which result in fair and equitable testing of limited English proficient students and that the testing of limited English proficient children will be a consideration in the development of all national educational programs.

2. Guidelines be developed to eliminate the unfair use of standardized testing in categorizing school populations. Item and population sampling may be investigated as possible interim solutions. This topic is one which could be the focus of another symposium.

3. Guidelines be developed to expand the use of alternative assessment practices in conjunction with standardized testing in the

program evaluation of federal education programs such as ESEA Title VII.

4. Incentives should be given for the development of computer technology for simulation testing which is an integral part of alternative methodologies.

My recommendations were addressed to the immediate concerns raised in this paper, however, it is important to realize that if we are to improve assessment practices for limited English proficient students we must begin simultaneous efforts in developing instrumentation in the child's native language, a topic which should also be the focus of a future symposium.