DOCUMENT RESUME

ED 367 684 TM 021 134

AUTHOR Rothman, Robert
TITLE Assessment Questions: Equity Answers. Proceedings of the 1993 CRESST Conference (Los Angeles, California, September 12-14, 1993). Evaluation Comment.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.; Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 94
CONTRACT R117G10027
NOTE 21p.
PUB TYPE Collected Works - Conference Proceedings (021)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Access to Education; *Conferences; Cost Effectiveness; Cultural Awareness; Cultural Differences; Definitions; *Educational Assessment; Educational Change; Educational Objectives; Elementary Secondary Education; *Equal Education; Outcomes of Education; Portfolios (Background Materials); Racial Differences; *Standards; *Test Bias; Test Construction; Testing; Testing Programs; Test Interpretation
IDENTIFIERS Alternative Assessment; *Authentic Assessment; Large Scale Programs; *Performance Based Evaluation; Reform Efforts

ABSTRACT

Focusing on one of the critical questions in the shift to new forms of assessment, researchers, policymakers, and teachers met at the 1993 Center for Research on Evaluation, Standards, and Student Testing (CRESST) to consider equity in assessment. The opening remarks of CRESST co-director Robert Linn stressed that equity is at the center of debates over assessment and new standards. Synopses of the remarks of a number of speakers are grouped under the following headings: (1) the definition of equity; (2) the evaluation of the fairness of assessments; (3) data from large-scale assessment programs; (4) costs of performance assessment; (5) equity and assessment design; (6) portfolios; (7) group assessment; (8) equity and the interpretation of assessment results; (9) equity and the research agenda; and (10) reports from working groups on the design of equity-sensitive performance assessments. The concluding remarks of Adam Urbanski of the Rochester (New York) Teachers Association pointed out that reforming schools is a long and difficult process that must involve the entire community. A list of CRESST and Center for the Study of Evaluation reports available is included. (SLD)

ED 367 684

UCLA's Center for the Study of Evaluation &
The National Center for Research on Evaluation, Standards, and Student Testing

EVALUATION COMMENT

Winter 1994

Assessment Questions: Equity Answers
Proceedings of the 1993 CRESST Conference

Robert Rothman, CRESST/UCLA

Zeroing in on one of the most critical issues in the shift to new forms of assessment, more than 200 researchers, policy makers, and teachers gathered at the UCLA campus September 12-14, 1993 to discuss questions surrounding equity.

Meeting at the 1993 CRESST conference, entitled "Assessment Questions: Equity Answers," participants outlined many of the concerns associated with the topic and offered some possible solutions.

CRESST Criteria of Fairness

CRESST Co-director Robert Linn said in opening remarks that equity is at the center of debates over standards and assessments. Fairness is one of the most critical of the criteria developed by CRESST to evaluate new assessments, he noted. At the same time, he said, the report of the National Council on Education Standards and Testing and the Clinton Administration's *Goals 2000* legislation have put equity at the top of the agenda in the federal and state governments.

"States and the national effort are focusing heavily on trying to establish ambitious content standards," Linn said. "And assessment has been central to all the work put forward in curriculum reform. This has led to demands for high standards of student performance assessed by new assessments congruent with the content standards."

"If you put the two together," Linn said, "that leads immediately to concerns about equity: what it means to give students a fair shot [at meeting the standards], especially if there are high stakes attached."

Dean Theodore R. Mitchell of the UCLA Graduate School of Education said the current debate represents a "unique historical moment." "For the first time," he said, "Americans are discussing both equity and excellence, taking into account both inputs and outcomes. We are at the threshold of a breakthrough," Dean Mitchell said.

But CRESST Associate Director Joan Herman cautioned that researchers do not yet have all the

See pages 13-16 for descriptions of eight new CRESST/CSE technical reports.

answers to enable us to walk through that door. "Most don't agree on what the central questions are," she said.

Defining Equity

While focusing their attention on the role of assessment, many agreed that the issue of equity involves the education system as a whole. "If we can adequately teach all youngsters, we don't need to be as concerned with equity," said Edmund W. Gordon, CRESST/City University of New York and Yale University.

"However," Gordon added, "that fact does not let those involved with assessments off the hook. Assessments themselves can be made more equitable," he said, "and assessments can help make inputs more equitable."

> To attain equity... opportunities for demonstrating competency must be diverse.

But while agreeing on equity as a goal, researchers offered differing perspectives on how to define the concept and how it might be achieved. To Gordon, equity is not equivalent to equality, but rather, sufficiency. To attain equity, resources must be distributed sufficiently and opportunities for demonstrating competency must be diverse. "You may expose all persons to the same standard," Gordon said, "but if the manner in which the standard is presented isn't one that matches the characteristics of each person, one cannot assert that it has been presented equitably."

Gordon also laid down a challenge to the assessment community and outlined three ways to make assessment more equitable.

First, make better use of the information provided from assessment to allocate resources equitably. Second, develop new instruments and procedures to tap students' affective traits, not just their cognitive skills. And third, conduct research and development work to build on what is known about pluralism. Although portfolios appear promising as a way of assessing students' abilities through diverse ways, Gordon warned against latching onto portfolios as the solution to every problem.

Gordon also cautioned that the problem of inequity is a problem of the larger society outside school. As long as society continues to reward winners and to screen out people "not like us," inequities will remain regardless of what happens to assessments.

Lauren B. Resnick, CRESST/LRDC, University of Pittsburgh, argued that assessment can play a major role in alleviating inequities. Like Gordon, Resnick said that the real issue is learning, and she said that changing assessment is vital to creating opportunity to learn for all Americans.

> "...equity...is the right to achieve at levels sufficient to participate productively, and in a rewarding way, economically and civically."

Resnick argued that we are heading down a *revolutionary* path, at the end of which all students will have a *right to achieve.* "What we mean by equity," she said, "is the right to achieve at levels sufficient to participate productively, and in a rewarding way, economically and civically."

Reaching such equity, she said, will require defining the level of achievement all children must attain, a process that is now under way through the development of national standards. In addition, it will require holding ourselves responsible for providing the op-

portunity for all students to achieve to the desired level. Although the *right* to achieve could eventually become a legal standard that would enable young people to demand the opportunities to achieve, in the meantime this *right* could serve as a moral obligation for society.

Lorraine McDonnell, CRESST/University of California, Santa Barbara, argued that attaining equity is a political, not a moral, challenge. She noted that Americans have long supported what she called *procedural equity*, or a process that ensures that everyone has access to valued goods. But *substantive equity*, or equal results, has never enjoyed public support, in part because it demands redistribution of resources. Achieving that type of equity, McDonnell argued, demands appealing to voters' self-interest for a better society in which young people are better educated. "One hopes for altruism, but it's hard to build a political majority that way," she said.

Similarly, Jeannie Oakes of the UCLA Graduate School of Education also doubted that moral suasion would be sufficient to ensure equity. Many schools, she said, lack the human and material resources needed to create the conditions to provide high levels of instruction for all students, particularly low-income, minority students. What is needed are opportunity-to-learn standards that would define a fair share of resources for schools. But that in itself may not be enough, Oakes added, because it is "astonishing how quickly good data disappear." Instead, she said, schools should make the data-collection process a part of the improvement process.

"My best guess at this point," Oakes said, "is that whatever process we engage in for the regular business of collecting information and feeding it back into our systems probably better look pretty much like the complex kind of teaching and learning we're hoping for in schools. Ideally, the process of assessing the quality of students' opportunities will become indistinguishable from the very effort to create and improve those opportunities."

Evaluating the Fairness of Assessments

Looking specifically at the issue of fairness and assessment, CRESST Co-director Robert Linn noted that the traditional methods of evaluating whether tests are fair—impact analyses, which measure differences in group performances; sensitivity reviews, which examine test content with an eye toward eliminating offensive or stereotyping material; and statistical analyses—are inadequate for use with performance assessments.

> Math assessments that demand substantial linguistic ability may be unfair to those lacking in that skill.

Linn proposed additional factors that must be considered to determine if new assessments are fair. These new factors include:

- The intent of the measure, or the extent to which assessments measure ancillary skills that might provide an advantage to a particular group. As an example, Linn noted that math assessments that demand a substantial amount of linguistic ability could prove a disadvantage to those with math skills who lack reading and writing ability.

- Comparability, or the extent to which the assessments allow variability in format or scoring.

- Choice of task. By allowing students to choose their tasks, such as reading a book, we may be providing an advantage to those who are already familiar with the books they choose.

- Delivery standards. To assure equity, one must look not only at assessments, but the provision of the instructional experiences to students.

Leigh Burstein, CRESST/UCLA, also said that examining students' opportunity to learn is essential to evaluate the fairness of new assessments. "You can't measure achievement without knowing the instructional conditions under which achievement occurs," he said. But he cautioned that studies of opportunities to learn have thus far only taken place in low-stakes environments, not in situations where schools were held accountable for providing such opportunities.

Burstein pointed out two ways in which students' opportunities could have an effect on their performance on alternative forms of assessments. First is the students' own experiences. Students accustomed to multiple-choice tests may balk at assessments that demand that they write responses, particularly in subjects other than English. He pointed out that in the 1990 National Assessment of Educational Progress, many students simply skipped over open-ended items, and there were ethnic group differences in the omit rates.

In addition to the variations in student experiences, differences in teachers' experiences can also affect students' performance, Burstein said. In reform environments, urban teachers tend to have fewer chances to participate in developing and scoring new assessments than do teachers from suburban and rural areas, since it costs districts to send teachers to such sessions.

Burstein described two studies designed to examine students' curricular experiences. The first looks at classroom "artifacts"—textbooks, logs of daily activities, homework assignments, in-class quizzes, and major assignments—and compares the findings with surveys of teachers that attempt to get at students' learning opportunities. The second, conducted as part of the California Learning Assessment System, asks students and teachers if they have done problems in their mathematics classes like the open-ended ones on the assessment.

> Performance-based assessments use a small number of tasks that may favor one group over another.

H.D. Hoover of the University of Iowa, however, said it is unclear whether differences in group performance on performance assessments reflect differences in curricular experiences or test bias. He said that, unlike multiple-choice tests, which can include a wide range of questions, performance-based assessments use a small number of tasks that may favor one group over another. A certain reading passage and its corresponding questions may appeal more to a student who has an inherent interest in the subject matter of that passage.

"For fairness, you need a diversity of content and contexts," said Hoover. "Those of us who build standardized tests—that's what we do. We ask lots of questions, and balance questions."

Data From Large-Scale Assessment Programs

Whether because of differences in opportunity to learn or because

of the assessments themselves, the gaps between advantaged and disadvantaged students are not closing as schools shift to new forms of assessment. In fact, researchers who have studied new large-scale assessment programs have found that the gaps may be widening.

Daniel Koretz, CRESST/RAND, said that some teachers in Vermont agree that that state's pioneering portfolio assessment program improved education for traditionally low-performing students. However, he noted, such improvements were erratic.

> The use of portfolio assessments requires teachers to adapt to everchanging curricular and instructional demands.

"We see sign after sign after sign that teachers vary enormously in response to performance assessment," Koretz said. "They vary in how quickly they adapt to demands, and to what is expected of them."

"Moreover," Koretz said, "the program itself of course has done nothing to alleviate the conditions that have plagued low-performing students, such as poverty." He concluded that improving the level of student performance through the use of portfolios is a "hard, arduous task" that will require money and an infrastructure to enable teachers to adapt to changing curricular and instructional demands.

Likewise, CRESST partner Mary Lee Smith from Arizona State University found that the first year of implementation of Arizona's performance-based assessment program failed to address the disparities between low-performing and high-performing schools. Although teachers had believed that the program would be a low-stakes exercise that would enable them to improve instruction, in part through developing and scoring the assessment, in practice the Arizona Student Assessment Program (ASAP), proved quite different. The scoring was done by a commercial publisher, not teachers, and the stakes went up when newspapers ranked school districts according to test scores. The state, moreover, provided little professional development to boost the capacities of schools.

Whether the ASAP program will eventually narrow the gaps in school performance is difficult to predict.

> We have seen teachers break an integrated unit into bits, and teach toward mastery of the bits.

"It is not yet clear," said Smith, "how ASAP will affect instruction, because or in spite of its high-stakes function, although there are already foreshadowings. It is predictable that districts with adequate resources will do what is necessary to raise low scores. Whether they will take the high road—undertaking the time-consuming and expensive professional and curriculum development work necessary to teach toward ambitious standards and a thinking curriculum—or the low road—finding the tricks to inflate scores—remains to be seen. At this point, we have already documented such activities as teachers focusing pupils' attention on those parts of the assessment that will be scored. We have also seen instances of what we call *dis-integrating*, in which teachers who lack a thorough understanding of constructivist teaching take what was designed to be an integrated unit, break it into bits, and teach toward mastery of the bits."

As with the large-scale pro-

grants, Lorrie A. Shepard, CRESST/University of Colorado at Boulder, said that in classroom assessments, gains in students' ability to perform well on performance tasks come slowly and unevenly. Drawing from her CRESST research involving third graders in three Colorado schools, Shepard cited a problem that asked the students to complete a table to determine how many pitchers (of 4 cups each) it would take to have enough cups for all the students in the class. There were no differences in performance between Anglo and Latino students, but initially very few students in either group could even attempt the problem. In the second year of the program, the majority of students could complete the table and wrote more to explain their answers than they did the first year, Shepard explained. But many still had a long way to go. "Anyone who thinks this can be put in place in a year or two is probably crazy," she said.

But Jennifer Harvey, a teacher at Cherry Drive Elementary School in Thornton, Colorado, who is part of the CRESST/University of Colorado at Boulder study, said even small gains are valuable. By using a variety of alternative assessment methods, Harvey now knows her students better than she ever did and she has evidence of their progress that she can show to parents. She also cited the case of one of her students, Jeff.

"I had given him a running record in October," said Harvey; "he read about a page and a half in 10 minutes. At the end of that time, I couldn't bear to watch his face anymore, it was so painful. So we stopped.... [He read] the exact same page in January. And he blew right through it. He did very well. Again, he was third grade, it was a preprimer, but that was progress. And I showed it to him and said, 'Look what you've done.' And he beamed and said, 'I'm getting better, aren't I?' That's where it's at, for me."

> Ideally, policy makers should weigh the costs and benefits of a proposed reform before implementation.

Costs of Performance Assessment

The benefits of performance assessment should also have a bearing on policy makers' decisions about whether to implement them, since policy makers should ideally weigh the costs and benefits of a proposed reform. But quantifying the potential benefits has proved elusive, noted David Monk of Cornell University, and as a result, most of the discussion of costs has focused solely on expenditures.

"You can't talk about cost without dealing with the benefit side of the equation," Monk said. "That's a problem with performance assessment. The simple fact is, we don't know very much about what performance assessment produces, or what kind of levels of resources are required for this to take place. In the absence of that kind of knowledge, you're in a bit of a dilemma trying to carry out a cost analysis that's more than an expenditure analysis."

Monk said that an analysis he conducted for the New Standards Project produced a range of cost estimates for performance assessments, depending on the extent to which the assessment is added on to existing programs and the extent to which every student is tested. In the *worst case* scenario, in which the assessment was an addition to existing programs and every student in three grades was tested, the assessment cost $97.4 million or $29 per pupil for a large state (Texas), $27.9 million

or $28 per pupil for a medium-sized state (Virginia), and $3.5 million or $37 per pupil for a small state (Vermont). In all cases, a little less than a fourth of the expenditures went toward staff development, and the total expenditures amounted to between 0.6 percent and 0.7 percent of each state's education budget.

Fritz Mulhauser of the U.S. General Accounting Office (GAO), which conducted an analysis of the cost of testing for Congress, said that the GAO study came up with an estimate for a proposed national test that was far lower than previous estimates—from $42 million for a multiple-choice test and little added time to $209 million for a test including short performance-based questions and 30 minutes of added testing time. But Mulhauser noted that the true cost depends on the purpose for such a test, and he urged Congress to be clear about the purpose before determining the proposed cost.

Similarly, Lawrence Picus, CRESST/University of Southern California, also raised a number of questions that need to be answered before determining whether the benefits of performance assessment exceed the costs. Among these questions are: whether we want to compare individual students or school districts, how many tasks are needed to provide a reliable estimate of student abilities, how much training is needed for teachers, and what are the competing claims for resources.

As in other issues of public policy, Picus said, deciding whether to invest in performance assessment involves tradeoffs: "How much of this do you want to do versus how much can you afford to do in terms of time and resources?"

Equity and Assessment Design

In looking ahead to new programs, researchers also discussed possible ways to ensure equity in assessment.

> Equity is one of the key criteria by which educators can judge tests.

One step toward that end is being taken by the National Council of Teachers of Mathematics (NCTM). In that group's *Assessment Standards for School Mathematics*, equity is one of the key criteria by which educators can judge tests, according to Thomas A. Romberg of the University of Wisconsin, Madison. Romberg said that the NCTM defines equity as providing all students the opportunity to demonstrate their mathematical power, and he noted that current tests do not match the council's goals.

> "...[Native Americans] also value patience, whereas tests demand rapid responses and immediate decisions."

Other researchers suggested that new assessments must take into account the needs of diverse students in order to be equitable. Michael Pavel of the UCLA Graduate School of Education said that Native American students are ill-served by traditional tests. Native Americans value placidity, a characteristic that may result in their being viewed as slow or backward, he said. Similarly, patience is valued, whereas tests demand rapid responses and immediate decisions.

He added that assessment should be better adapted to Native American people who need to improve on their academic performance and make teachers more

aware of how they can address these academic needs. Therefore, to assist Native Americans, assessment results should be used to guide student learning, refine curriculum, and improve instruction.

Charlene Rivera of George Washington University said assessment reform must also consider students for whom English is not the native language. In addition to learning content knowledge and skills, she pointed out, "English language learners" (ELLS) are also learning a second language, something native English speakers do not have to do.

"To date, reform efforts have not considered needs of ELLS students," Rivera said. "The guiding assumptions have been: What will work for monolingual students will also work for ELLS students. Once ELLS students learn a little English, new improved assessment systems will fit them too. However, experience doesn't support this assumption. While ELLS students can and do learn to high standards, assessing their achievements in the same way as their monolingual peers will greatly underestimate their accomplishments and potential."

Rivera added that schools currently include English language learners in inappropriate testing programs or else exempt them from tests altogether. "I am on the side of trying to develop a middle ground where there is accountability for student learning," she said. "If you exclude them completely, they are not considered in the policy debate. But how to assess them? There is no definitive answer. The best practices will result from experimentation."

> Experiments already underway suggest possible solutions for traditionally underserved groups.

Some experiments already under way suggest possible solutions for English language learners and other traditionally underserved groups. One such experiment is a primary-grades assessment currently being developed by a consortium of six states under the auspices of the Council of Chief State School Officers. Jackie Cheong of the University of California, Davis, said one of the guiding principles of that effort is to provide multiple assessment strategies to tap a range of what students from diverse backgrounds know and are able to do. Establishing the link between classroom assessments, such as portfolios, and accountability is one way to further the use of multiple assessments.

Portfolios

Likewise, portfolios can also provide diverse opportunities for students to demonstrate their skills and knowledge. A school project in Pasadena, California, that integrates language arts and visual and performing arts instruction by asking students to create works of art and to reflect on their own and other works has shown dramatic results among a group of formerly low-achieving children, according to Pam Aschbacher, CRESST/UCLA.

But Aschbacher also pointed out that implementing portfolios alone may not improve educational opportunities for all youths and the portfolios themselves can demonstrate this. In a separate project, Aschbacher examined portfolios to open a window on classroom instruction and found that teachers were not always meeting reform objectives. Under the program, known as *Humanitas*, students were expected to be personally invested in the work that they're doing in class, to make interdisciplinary connections, to engage in com-

plex reasoning, and to evaluate themselves and show some growth over time, among other goals.

The portfolios showed a direct relation between the kind of assignments teachers gave their students and what students learned. For example, she found that students' work showed more higher order thinking skills and interdisciplinary connections when their class assignments *required* them to make those connections and to use complex thinking. "That's a strong message," said Aschbacher, "to these teachers that says: 'Look, when you use assignments that do not explicitly ask kids to do the kind of thinking this program calls for, they don't do it. You get what you ask for.'"

Similarly, Catherine Smith of the Michigan Department of Education said that a program in that state to use portfolios to gauge students' workforce readiness also raised equity concerns. "Students in inner-city and rural areas were less likely than those in suburbs to be aware that they could use evidence from part-time jobs and other afterschool activities to demonstrate readiness skills," she said.

Maryl Gearhart, CRESST/UCLA, also reported that implementing portfolio assessment alone may not have much impact on classroom practice. Describing two R&D projects on classroom assessment, one focused on writing and the other on mathematics, Gearhart said, "I have found that elementary teachers do not typically develop in their understandings of student work without a substantive focus on subject matter. New assessments require teachers to make informed judgments," she said, "but teachers cannot judge work that they do not understand. In our projects, the goal is to create prototypes of assessment practices *integrated* with reform curricula. We can't simply exhort teachers to collect student work and assess it," said Gearhart. "We need to give them specific models built upon specific curriculum."

Group Assessment

Similar caution flags went up over another potential avenue for ensuring equity: the use of group work. Wayne Neuberger of the Oregon Department of Education suggested that allowing students to work in groups in advance of an assessment could *level the playing field* by enabling students to compare notes and ensure that they all had the same background to prepare for the assessment. The New Standards Project, of which Oregon is a member, provides an opportunity to do just that.

But Noreen Webb, CRESST/UCLA, noted that group interactions are complex and that simply asking students to cooperate may not result in everyone's acquiring shared knowledge and understandings. Factors such as the composition of the group and the incentives they have for working together can influence whether group interactions are functional or dysfunctional, Webb said.

Moreover, she added, relying on group assessments as a measure of student abilities may adversely affect equity. Depending on the way groups work together, she said, group performances may not be a valid measure of the performance of individual members, and, as a result, group performances may mask individual deficiencies. If that is the case, Webb warned, low performers may miss out on needed instructional help.

Equity and Interpreting Assessment Results

In addition to the design of new assessments, the way results are interpreted also has a bearing on questions of equity. David Bayless of Bayless and Associates

said that the traditional way of analyzing assessment results is to look at subgroups, such as gender, socioeconomic status, and ethnicity. But those analyses can provide little guidance for improvement, because they cannot be changed. "It's difficult to change [a student's] gender," he said. Rather than select demographic subgroups that are easy to measure, Bayless argued, analysts should examine assessment results according to factors that can be influenced by intervention, such as opportunity to learn.

Examining single test scores may mask important differences in subgroups' performance.

Bengt Muthén, CRESST/UCLA, said that examining single test scores may mask important differences in subgroups' performance. In one study, Muthén said he is using multivariate analysis to gain a broader picture of student performance on the National Assessment of Educational Progress. Preliminary findings suggest that while female and male students perform about equally well overall, male students outperform females on certain subscales, such as measurement. Muthén is also analyzing the results of the Longitudinal Study of American Youth to detect the effects of tracking and coursetaking on student performance over time.

Looking at a separate way of examining possible sources of bias, Jamal Abedi of CRESST/UCLA said that different methods of analyzing interrater reliability may yield different estimates. As a result, he suggested using different applicable approaches. "I cannot name a single best approach to establishing interrater reliability," Abedi said. "Depending on the form of data, compute as many applicable approaches as you can, and then draw your conclusions based on those outcomes."

Michael T. Nettles of the University of Michigan proposed an additional step. After citing ethnic group differences in performance on various assessments, he said that we should look at assessments that show good results with traditionally low-performing students.

Under the auspices of the Ford Foundation, Nettles said that during the next year he expects to develop a second symposium on equity and educational testing and assessment and develop a program to identify exemplary assessment programs. One of the criteria for being considered exemplary should be that the assessments have demonstrated making a contribution to improving the success of poor and minority student populations.

Equity and the Research Agenda

Although some answers to the equity questions surrounding new forms of assessment are beginning to emerge, the research agenda remains long. Pauline Brooks, CRESST/UCLA, outlined a host of questions in light of the unequal representation of gender, socioeconomic status, and diverse cultures throughout the history of testing and assessment. Included on her list were:

- the effects of the teacher's cultural expectations on his/her judgments of culturally diverse student performances;
- the extent to which the new assessments provide varying opportunities for students of different cultural/SES/language backgrounds to demonstrate their knowledge;
- the correlation between

performance on standardized achievement tests and current performance assessments—do the racial/cultural and SES gaps narrow, remain about the same, or widen with the use of performance assessments?

- the characteristics of teachers' interactions with different cultural groups of students, in both instructional and assessment settings; and

- relative levels of support for the new assessments among communities that vary economically and culturally.

Working Group Results

In addition, working groups representing various key constituencies outlined action plans for designing equity-sensitive performance assessments. The groups recommended the following:

Policy Makers:

- Clearly articulate the purposes of new assessments, so that the public understands that assessments are aimed at meaningful and effective accountability, not an attempt to evade accountability.

- Make sure assessment for "classroom utility" and assessment for public accountability are in sync, so that teachers have an incentive to focus on improved instruction.

- Keep expectations high for all students to eliminate any incentive to relegate low-performing students to a second-class education.

Practitioners:

- Make the purposes and learning outcomes of new assessments clear and make sure that large-scale and classroom-level assessments are integrated.

- Encourage flexibility so that all students have the time and opportunities to succeed on new assessments.

- Provide choices in assessment alternatives to children that are appropriate to their ethnicity, their gender, the possible existence of handicap, the language that they use, etc. Incorporate diverse groups during the design and piloting process so that we engage all communities and constituencies.

- Develop university-school partnerships to share knowledge about children's learning and provide ongoing staff development for teachers.

- Conduct longitudinal and comparative studies to examine the impact of new assessment strategies on teachers and students.

Foundations:

- Make sure new programs are practical and feasible so that they can be implemented in schools and yield useful findings.

- Make sure programs are credible to parents, teachers, and funding agencies and that the programs are actually implemented.

- Try to focus on the "hard-to-crack" cases.

Business Community:

- Educators, businesses, and the community at large need to find a better way to collaborate to produce

diagnostic assessments that enable our students to advance their own goals in the classroom and in the workplace.

- Assessments must also be predictive and show students' readiness for the workplace as well as their ability to transfer skills from school to work.

- Consider whether all assessments must be administered in school or by teachers or whether other approaches are possible, depending on the purpose.

Media:

- Highlight the standards and content of new assessments as a way of showing what all students are expected to learn.

- Report opportunity to learn data and classroom environments to provide a better understanding of equity.

- Develop links between researchers and the media to provide context for test-score data.

Parents:

- Involve parents in all aspects of the development of new assessments via advocates or site-based management, helping to ensure that they are fair and better for all children.

- Collect data on instruction, testing participation, growth in student performance, and teacher grading, to accompany test-score data.

- Conduct research on parent involvement, *toxic schools*, the role of churches, and communication of testing information.

In Conclusion

In concluding remarks, Adam Urbanski, the president of the Rochester (NY) Teachers Association, pointed out that reforming schools so that all students learn at high levels is a long and difficult process that must involve the entire community. But he said that setting standards and developing new assessments is the essential starting point of such efforts in order to ensure both excellence and equity.

"I respectfully suggest that all students are learning already," said Urbanski. "They're just not learning the same things. Some students are learning math and English and foreign languages and the arts and physics. And some are learning a lot about exclusion and failure and discrimination and lack of opportunities and low expectations of them. But you can't stop students from learning. You can only assist in channeling them to certain 'kinds' of learnings. That is why the whole issue of standards and assessment is so germane. They are absolutely essential and not only essential, but indeed the necessary starting point for all other reforms. School change need not be a choice between making things better or making things fair. We're capable of doing both. We must make things better and be fair in doing it."

Robert Rothman is a visiting researcher at CRESST/UCLA.

CRESST/CSE Technical Reports

The following assessment reports are now available by calling (310) 206-1532. Or you may fill in the order form on page 20 and mail to CSE/CRESST Annex, Graduate School of Education, 405 Hilgard Ave., Los Angeles, CA 90024-4108.

VERMONT UPDATE

Can Portfolios Assess Student Performance and Influence Instruction? The 1991-92 Vermont Experience

Daniel Koretz, Brian Stecher, Stephen Klein, Daniel McCaffrey, and Edward Deibert, RAND

CSE Technical Report 371, 1993 ($9.00)

Vermont's statewide assessment initiative program has garnered widespread attention nationwide because of its reliance on portfolios of student work. This 145-page CRESST/RAND report describes results of a multifaceted evaluation of the program and provides information about implementation of the Vermont assessment, program effects on educational practice, reliability and validity of portfolio scores, and tensions that exist between assessment and instructional reform.

"Findings from the evaluation," said the research team, "suggest that the assessment program resulted in changes in curriculum content and instructional style." Additionally, the researchers noted that the amount of classroom time devoted to problem solving increased, as did the amount of time students worked in small groups. Finally, portfolios seem to increase teachers' enthusiasm for their subjects and for teaching.

While there was widespread support for the reform at the school level throughout the state—nearly one-half of the schools were voluntarily expanding the use of portfolios to other grade levels—substantial problems remain. The mathematics portfolio assessment created new burdens for principals, teachers and students, including demands on teachers' time and school resources. Over 80% of fourth-grade teachers and over 60% of eighth-grade teachers reported that they often had difficulty covering the required curriculum. Researchers anticipate that in time some of these demands are likely to decline, although others represent continuing burdens.

"The Vermont experience has important implications for reforms that are underway or under consideration in other jurisdictions," said the researchers, "but only time and careful scrutiny will show how fully the goals of the program—and of similar reform programs centered on performance assessment—can be met."

MORE ABOUT VERMONT

Interim Report: The Reliability of Vermont Portfolio Scores in the 1992-93 School Year

Daniel Koretz, Stephen Klein, Daniel McCaffrey, and Brian Stecher, RAND

CSE Technical Report 370, 1993 ($3.00)

This interim report provides the first results of the second year of the Vermont assessment program, focusing on program implementation, effects on education, and the quality of performance data.

"The program was altered in many ways in 1992-93," said the CRESST/RAND research team, "which resulted in a clear increase in the reliability with which mathematics portfolios were scored." However, the researchers added that while this progress is encouraging, scoring reliability in mathematics needs to be increased further if the program goals are to be achieved. Refining or simplifying scoring rubrics and placing further restrictions on types of tasks

NEW REPORT

CRESST/CSE Technical Reports

considered acceptable for inclusion in mathematics portfolios are among the types of clarifications that may result in increased reliability.

"In contrast," said the researchers, "the reliability of writing portfolio scores did not improve substantially and was considerably lower than in mathematics." The researchers believe that it is unrealistic to expect a substantial rate of improvement in the reliability of the writing portfolio scores unless the program itself is substantially revised.

☆ Comparability Across Assessments: Lessons From the Use of Moderation Procedures in England

Elizabeth Burton and Robert L. Linn

CSE Technical Report 369, 1994 ($4.00)

Although there is considerable interest in developing a system of performance-based examinations in the United States, there is a general lack of agreement on how to compare the results of different performance assessments to sets of common national standards. This paper addresses the "comparison" problem, drawing on two major approaches used in England, *moderation by inspection* and *statistical moderation*, to link performance assessments to sets of common standards.

Although the United States will probably not develop a program of assessments exactly like those used in England, it is likely that the procedures used to compare the assessments will be similar.

"Currently English secondary school exams in various subjects are developed and administered by nine examination boards," said Burton and Linn. "Individual schools are free to choose the examination board that best fits their standards. While local control and high quality of assessments are maintained, the comparison of scores across the boards is problematic," added the researchers.

In this report, Burton and Linn discuss the advantages and problems of *moderation by inspection* and *statistical moderation*, together with an explanation of why neither approach is satisfactory by itself. The authors concluded that some combination of the two approaches may be necessary. "Neither a pure moderation by inspection nor a strict statistical moderation system is likely to meet this [link between assessments and standards] need," said Burton and Linn. "It seems more likely that some sort of hybrid system will be required..."

☆ Results From the New Standards Project Big Sky Scoring Conference

Lauren Resnick, Daniel Resnick, and Lizanne DeStefano

CSE Technical Report 368, 1993 ($3.50)

Partially funded by CRESST, the New Standards Project is an effort to create a state- and district-based assessment and professional development system that will serve as a catalyst for major educational reform. In 1992, as part of a professional development strategy tied to assessment, 114 teachers, curriculum supervisors, and assessment directors met to score student responses from a field test of mathematics and English language arts assessment. The results of that meeting, the Big Sky Scoring Conference, were used to analyze for comparability across holistic and anaholistic scoring methods.

"Interscorer reliability estimates," said the researchers, "for reading and writing were in the moderate range, below levels achieved with the use of large-scale writing assessment or standardized tasks. Low reliability limits the use of [the] 1992 reading and writing scores for making

CRESST/CSE Technical Reports

judgments about student performance or educational programs," concluded the research team.

However, interscorer reliability estimates for math tasks were somewhat higher than for literacy. For six out of seven math tasks, reliability coefficients approached or exceeded acceptable levels.

The findings suggest that the large number and varied nature of participants may have jeopardized the production of valid and reliable data. "Scorers reported feeling overwhelmed and overworked after four days of training and scoring," said the researchers.

Despite these difficulties, evidence was provided that reliable scoring of large-scale performance assessments can be achieved when ample time is provided for training, evaluation, feedback, and discussion; clear definitions are given of performance levels and the distinctions between them; and well-chosen exemplars are used.

☆ Parent Opinions About Standardized Tests, Teacher's Information and Performance Assessments

Lorrie A. Shepard and Carribeth L. Bliem

CSE Technical Report 367, 1993 ($4.00)

Using parents of third-grade students in a working-class and lower-middle-class school district, researchers ... this study set forth to ascertain parents' opinions about assessment, including their opinions about standardized tests versus performance assessments. The researchers sought answers to several questions including "How do parents in the sample respond to Gallup Poll questions about the desirability of standardized national tests and the potential uses for standardized test results?" As part of the study, parents were given an opportunity to review performance assessment tasks and decide what type of assessment, standardized or performance, was most suitable for classroom use.

The results indicated that when allowed to look closely at performance assessment problems, most parents endorsed performance assessments for district purposes and especially preferred their use in classroom contexts.

☆ Teachers' Ideas and Practices About Assessment and Instruction

Hilda Borko, Maurene Flory, and Kate Cumbo

CSE Technical Report 366, 1993 ($4.00)

Participants involved in this study were part of a year-long intervention designed to help teachers develop performance assessments in reading and mathematics. Seeking to evaluate teachers' knowledge, beliefs, and practices about assessment and instruction, the researchers also studied the changes that occurred to teachers during the first semester of the intervention program.

Findings from the study indicated that the performance assessment development and implementation process resulted in teachers having better understandings and new insights into students' thinking and learning than when teachers relied exclusively on more traditional forms of assessment. However, it was not clear to what extent teachers changed their instructional programs to take advantage of their newly gained insights. Based on their observations so far, researchers feel confident that as the program continues, more extensive changes will occur.

☆ Dilemmas and Issues in Implementing Classroom-Based Assessments for Literacy

Elfrieda H. Hiebert and Kathryn Davinroy

CSE Technical Report 365, 1993 ($3.50)

Researchers in this study in-

CRESST/CSE Technical Reports

vited third-grade teachers from an urban school district to collaborate in a classroom-based literacy assessment project. The study focused on a series of literacy workshops designed to implement a long-standing perspective on curriculum, instruction and assessment adapted to classroom-based assessment.

Some of the early outcomes from observations and transcriptions of the workshops indicated that teachers struggled with a variety of issues including the task of embedding assessments like running records and written summaries into their instructional programs. Despite many challenges, at least one of the schools moved quickly to implement the assessments and use the information the assessments provided.

Dilemmas and Issues for Teachers Developing Performance Assessments in Mathematics

Roberta J. Flexer and Eileen A. Gerstner

CSE Technical Report 364, 1993 ($4.00)

This paper examines some of the dilemmas and issues that arose during the first two terms of work with teachers participating in the development of assessments in mathematics, and reports on changes in their instruction and assessment as a result of the project.

During the study, many dilemmas and issues arose that were unique to each of the three schools studied, but the most challenging problem was teachers' focus on what was important to teach (and therefore assess), and how children could learn what was taught—all within the constraints of limited teacher time. As expected, preliminary results of the project were mixed, but hopeful. Researchers believe that future development and implementation of performance assessments in these classrooms hinge on teachers' beliefs in these assessments as useful and practical tools.

***Whose Work Is It?* A Question for the Validity of Large-Scale Portfolio Assessment**

Maryl Gearhart, Joan L. Herman, Eva L. Baker, and Andrea K. Whittaker

CSE Technical Report 363, 1993 ($3.00)

Portfolio assessment represents a growing commitment to bridge the worlds of public accountability and private classroom, and policy maker and child. Thus, within the move toward further authenticity, portfolios support performance-based assessments that may incorporate shared readings of common background texts, collaborative planning, and opportunities for students to revise their work.

Based on an in-depth analysis of nine elementary school teachers actively using writing portfolios in their classrooms, the researchers of this study focused on a technical issue not yet directly investigated in R&D studies of portfolio assessment: *"Whose work is it?"* If students collaborate with peers or receive assistance from parents or teachers, the authorship of classroom work of the student is in question. Focusing here just on the teachers' contributions to student work, the authors documented patterns of instructional support across writing assignments and students. The work raises technical issues concerning the meaningfulness of 'student' scores derived from assessment of student portfolios.

More CRESST/CSE Technical Reports

Performance-Based Assessment and What Teachers Need
Higuchi
CSE Technical Report 362, 1993
($4.00)

Sampling Variability of Performance Assessments
Shavelson, Gao, & Baxter
CSE Technical Report 361, 1993
($4.00)

Raising the Stakes of Test Administration: The Impact on Student Performance on NAEP
Kiplinger & Linn
CSE Technical Report 360, 1993
($4.00)

Issues in Innovative Assessment for Classroom Practice: Barriers and Facilitators
Aschbacher
CSE Technical Report 359, 1993
($4.50)

Writing What You Read: Assessment as a Learning Event
Wolf & Gearhart
CSE Technical Report 358, 1993
($4.00)

Omitted and Not-Reached Items in Mathematics in the 1990 National Assessment of Educational Progress
Koretz, Lewis, Skewes-Cox, & Burstein
CSE Technical Report 357, 1992,
($4.00)

Latent Variable Modeling of Growth with Missing Data & Multilevel Data
Muthén
CSE Technical Report 356, 1992
($2.50)

The Reliability of Scores From the 1992 Vermont Portfolio Assessment Program
Koretz, Stecher, & Deibert
CSE Technical Report 355, 1993
($3.00)

Assessment of Conative Constructs for Educational Research and Evaluation: A Catalogue
Snow & Jackson
CSE Technical Report 354, 1992
($8.00)

The Apple Classrooms of Tomorrow[sm]: The UCLA Evaluation Studies
Baker, Gearhart, & Herman
CSE Technical Report 353, 1993,
($3.50)

Collaborative Group Versus Individual Assessment in Mathematics: Group Processes and Outcomes
Webb
CSE Technical Report 352, 1993,
($4.00)

Educational Assessment: Expanded Expectations and Challenges (1992 Thorndike Award Address)
Linn
CSE Technical Report 351, 1992,
($3.50)

The Vermont Portfolio Assessment Program: Interim Report on Implementation and Impact, 1991-1992 School Year
Koretz
CSE Technical Report 350, 1992
($6.00)

Design Characteristics of Science Performance Assessments
Glaser, Raghavan, & Baxter
CSE Technical Report 349, 1992
($3.00)

Accountability and Alternative Assessment
Herman
CSE Technical Report 348, 1992
($4.00)

Benchmarking Text Understanding Systems to Human Performance: An Exploration
Butler, Baker, Falk, Herl, Jang, & Mutch
CSE Technical Report 347, 1991
($5.00)

The Influence of Problem Context on Mathematics Performance
Webb & Yasui
CSE Technical Report 346, 1992
($4.00)

Report on Multilevel and Longitudinal Psychometric Models: Latent Variable Models for Analysis of Growth
Muthén & Nelson
CSE Technical Report 345, 1992
($2.50)

Measurement of Workforce Readiness Competencies: Design of Prototype Measures
O'Neil, Jr., Allred, & Baker
CSE Technical Report 344, 1992
($4.00)

Measurement of Workforce Readiness: Review of Theoretical Frameworks
O'Neil, Jr., Allred, & Baker
CSE Technical Report 343, 1992
($4.00)

CRESST/CSE Technical Reports

Will National Tests Improve Student Learning?
Shepard
CSE Technical Report 342, 1991
($3.00)

Implications for Diversity in Human Characteristics for Authentic Assessment
Gordon
CSE Technical Report 341, 1991
($2.00)

The Natural Language Sourcebook
Read, Dyer, Baker, Mutch, Butler, Quilici, & Reeves
CSE Technical Report 340, 1991
($15.00)

Language Assessment Instruments: LAUSD Language Development Program for African American Students
Butler, Herman, & Yamaguchi
CSE Technical Report 339, 1991
($4.00)

Discovering What Schools Really Teach: Designing Improved Indicators
McDonnell, Burstein, Ormseth, Catterall, & Moody
CSE Technical Report 338, 1990
($5.00)

Writing Portfolios at the Elementary Level: A Study of Methods for Writing Assessment
Gearhart, Herman, Baker, & Whittaker
CSE Technical Report 337, 1992
($4.00)

A New Mirror for the Classroom: A Technology-Based Tool for Documenting the Impact of Technology on Instruction
Gearhart, Herman, Baker, Novak, & Whittaker
CSE Technical Report 336, 1992
($5.00)

Cross-State Comparability of Judgements of Student Writing: Results From the New Standards Project
Linn, Kiplinger, Chapman, & LeMahieu
CSE Technical Report 335, 1992
($5.50)

Effects of Standardized Testing on Teachers and Learning—Another Look
Herman & Golan
CSE Technical Report 334, 1991
($5.50)

Conceptual Considerations in Instructionally Sensitive Assessment
Burstein
CSE Technical Report 333, 1990
($2.00)

Multilevel Factor Analysis of Class and Student Achievement Components
Muthén
CSE Technical Report 332, 1990
($3.00)

Complex, Performance-Based Assessment: Expectations and Validation Criteria
Linn, Baker, & Dunbar
CSE Technical Report 331, 1991
($3.00)

The Validity and Credibility of the Achievement Levels for the 1990 NAEP in Mathematics
Linn, Koretz, Baker, & Burstein
CSE Technical Report 330, 1991
($6.00)

For a complete list of more than 140 technical reports, monographs, and resource papers, please call Kim Hurst at (310) 206-1532.

UCLA's Center for the Study of Evaluation &
The National Center for Research on Evaluation, Standards, and Student Testing
Eva L. Baker, Co-director
Robert L. Linn, Co-director
Joan L. Herman, Associate Director
Ronald Dietel, Editor
Katharine Fry, Editorial Assistant
Brenda R. Thomas, Layout

The work reported in this publication was supported under the Educational Research and Development Center Program cooperative agreement number R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this publication do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

MONOGRAPHS AND RESOURCE PAPERS

MONOGRAPHS

Assessing Student Achievement: A Profile of Classroom Practices
Dorr-Bremme & Herman
CSE Monograph 11, 1986
($11.00)

Evaluation in School Districts: Organizational Perspectives
Bank & Williams (Editors)
CSE Monograph 10, 1981 ($7.50)

Values, Inquiry and Education
Gideonse, Koff, & Schwab (Editors)
CSE Monograph 9, 1980 ($11.00)

Toward a Methodology of Naturalistic Inquiry in Educational Evaluation
Guba
CSE Monograph 8, 1978 ($4.50)

The Logic of Evaluative Argument
House
CSE Monograph 7, 1977 ($4.50)

Achievement Test Items—Methods of Study
Harris, Pearlman, & Wilcox
CSE Monograph 6, 1977 ($4.50)

RESOURCE PAPERS

Writing What You Read: A Guidebook for the Assessment of Children's Narratives
Wolf & Gearhart
CSE Resource Paper 10 ($4.00)

Improving Large-Scale Assessment
Aschbacher, Baker, & Herman
CSE Resource Paper 9 ($10.00)

Improving Opportunities for Underachieving Minority Students: A Planning Guide for Community Action
Bain & Herman
CSE Resource Paper 8 ($11.00)

Designing and Evaluating Language Programs for African-American Dialect Speakers: Some Guidelines for Educators
Brooks
CSE Resource Paper 7 ($2.00)

A Practical Approach to Local Test Development
Burry, Herman, & Baker
CSE Resource Paper 6 ($3.50)

Analytic Scales for Assessing Students' Expository and Narrative Writing Skills
Quellmalz & Burry
CSE Resource Paper 5 ($3.00)

Criteria for Reviewing District Competency Tests
Herman
CSE Resource Paper 4 ($2.00)

Issues in Achievement Testing
Baker
CSE Resource Paper 3 ($2.50)

Evaluation and Documentation: Making Them Work Together
Burry
CSE Resource Paper 2 ($2.50)

An Introduction to Assessment and Design in Bilingual Education
Burry
CSE Resource Paper 1 ($3.00)

FOLD AND SECURE

Place Postage Here

UCLA
CSE/CRESST Annex
Graduate School of Education
405 Hilgard Avenue
Los Angeles, California 90024-4108
Mail Code: 139448

Order Form

Attach additional sheet if more room is needed. Form is pre-addressed on reverse.

CSE Reports/Monographs/Resource Papers/Videotapes

Report Number	Title	Number of copies	Price per copy	Total Price

POSTAGE & HANDLING
(Special 4th Class Book Rate)

Subtotal of	$0 to $10	add $1.50
	$10 to $20	add $2.50
	$20 to $50	add $3.50
	over $50	add 10% of Subtotal

ORDER SUBTOTAL _____

POSTAGE & HANDLING (scale at left) _____

California residents add 8.25% _____

TOTAL _____

Orders of less than $10.00 must be prepaid

Your name & mailing address—please print or type:

_____

_____

_____

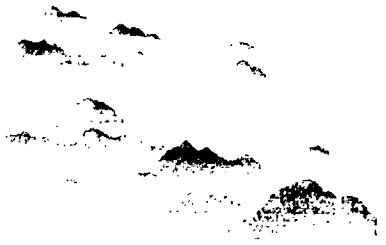
❑ Payment enclosed ❑ Please bill me

❑ I would like to receive free copies of the *CRESST Line* and *Evaluation Comment* publications.

UCLA
CSE/CRESST Annex
Graduate School of Education
405 Hilgard Avenue
Los Angeles, California 90024-4108
Mail Code: 139448

NONPROFIT ORG.
U.S. POSTAGE
PAID
U.C.L.A.

ADDRESS CORRECTION REQUESTED (EF 13)



Librarian
ERIC
UCLA
405 Hilgard, 96 Powell Library
Los Angeles CA 90024