C-R-E-S-S-T

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

Technical Report

You can view this document on your screen or print a copy.

- UCLA Center for the Study of Evaluation

in collaboration with:

- University of Colorado
- NORC, University of Chicago
- LRDC, University of Pittsburgh
- The RAND Corporation

Performance-Based Assessment for Accountability Purposes: Taking the Plunge and Assessing the Consequences

CSE Technical Report 390

Leigh Burstein
CRESST/University of California, Los Angeles

November 1994

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported herein was supported in part under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

Leigh Burstein passed away on July 7, 1994. In honor of his memory, we are publishing this report with virtually no editorial changes. We ask anyone who references or quotes from this paper to note that Leigh did not review the final publication. It is possible that Leigh would have made significant changes, or perhaps none at all.

Ron Dietel
CRESST Director of Communications

PERFORMANCE-BASED ASSESSMENT FOR ACCOUNTABILITY PURPOSES: TAKING THE PLUNGE AND ASSESSING THE CONSEQUENCES[1]

Leigh Burstein
CRESST/University of California, Los Angeles

At the 1990 Education Commission of the States (ECS) Assessment Conference in Boulder, the looming rapid expansion of interest in performance assessments was obvious to almost everyone. From a few sessions featuring presentations by the super-advocates at earlier conferences, the 1990 conference featured sessions by some of the major mainstream scholars expounding their views on performance assessment while others were voicing skepticism along classical traditional lines. As best I could tell, the commercial test publishers were nervous about when and how to risk stepping into these virtually uncharted waters in order to avoid losing out on new state assessment contracts and the like. Against this backdrop, Joy Frechtling organized a 1991 AERA symposium on performance assessment for accountability purposes to reflect on what had transpired at the ECS conference and on what the future for assessment might be.

---

[1] This paper was originally presented as part of an invited symposium, *Performance Assessment and Accountability Programs: Match or Mismatch?*, at the annual meeting of the American Educational Research Association, Chicago, April 4, 1991. An earlier version was also presented at the Iowa Policy Seminar organized by the North Central Regional Educational Laboratory in Des Moines, Iowa, October 18, 1990.

Joy's timing was prescient. The press for alternative assessment has exploded. To the policy and assessment communities, alternative assessments are no longer fringe activities practiced by curriculum and cognitive psychology visionaries. Everybody is doing it, or will be in some form soon. The 1991 NCME annual meeting program demonstrated just how far the traditional measurement community has come in exploring technical and policy issues in performance assessment. The 1991 ECS conference considered virtually nothing else. Remarks that I made at Boulder in 1990 and in talks in Iowa and Washington that would have been viewed as radical and perhaps heresy within the measurement community a short time ago are definitely mainstream now. So rather than representing the radical fringe of the measurement community, I may find myself in the uncomfortable position of being firmly in the middle.

My misgivings about being viewed as mainstream notwithstanding, my purpose is to comment on a few of the issues in alternative assessment for accountability purposes as I see them. In doing so I will rely heavily on my talks with folks in California, Iowa and Washington; on the ideas of my colleagues at the Center for Research on Evaluation, Standards, and Student Testing (CRESST), especially those contributed largely by Eva Baker and Bob Linn; and on a few of my recent ventures into the performance assessment world. I am now an active, though neophyte, practitioner of this new art. Rich Shavelson, Ed Haertel, Don Barfield, and I plan to collaborate with Cathy Comfort of the California Assessment Program (CAP) to develop a strategy for generating alternative hands-on assessment tasks in science (Shavelson, Comfort, Barfield, Burstein, & Haertel, 1991). I continue to participate in the design of the new CAP with its heavy "authentic" assessment components (California Assessment Policy Committee, 1991).[2] I am also working with other CRESST colleagues on various cross-cutting technical issues and collaborations with states in developing and understanding the technology of alternative assessments over the next few years. It will undoubtedly be exciting, frustrating, and exhausting; we hope it will be important and useful vis-à-vis the improvement of educational policy and practice as well as educational research.

[2] Editor's note: The new CAP was eventually named the California Learning Assessment System (CLAS).

Definitional Problems

Alternative assessment, performance assessment, authentic assessment, portfolios, mini-investigations, and writing samples are tied together in most broad discussions of new forms of performance-based assessment. This umbrella incorporates virtually all production oriented assessment tasks of more than, say, two sentences. But all alternative forms of assessment do not have the same attributes in terms of technical and feasibility criteria.

There is also a definitional problem with the term "accountability." Student accountability, that is, decisions about passing, promotion, selection, and placement, represent one set of problems. School, program, and teacher accountability reflect another set of conditions. The major features of difference are the implications for the precision of estimation of individual performance and the associated time demand (and resources) necessary for high precision. There are also huge differences in consequences when these new assessments are viewed as part of accountability at the student level (some form of stakes for the student through certification of some sort) or at higher levels. Given current trends (e.g., discussions of national examinations such as America 2000 [U.S. Department of Education, 1991], the Resnick-Tucker proposal (Learning Research and Development Center & National Center on Education and the Economy, 1990], the National Education Goals Panel; the Secretary's Commission on the Achieving Necessary Skills [U.S. Department of Labor, 1991]; the Special Study Panel on Education Indicators [1991]), however, we can anticipate that even those states and districts that avoided getting into the business of providing individual student scores will have to now (it will definitely happen in California). Thus, some of the ways of glossing over certain technical and feasibility considerations in performance assessment will not carry as much weight when civil rights and legal considerations with respect to individual students become the norm by which the viability of alternative assessments must be judged.

Tradeoffs in Validity of Inference

If assessments are intended to determine what students know and can do (or what they have been taught or have learned), then the typical, standardized multiple-choice (MC) tests involve a tradeoff that has been largely ignored up to now in existing accountability assessments. Traditional technical criteria ensure

highly reliable measurement of a subset of the attribute of interest; namely, the specific skills assessed on average in the items administered, along with associated skills at responding to particular types of questions administered under a particular set of conditions. That is, ability to respond to substantive probes via MC testing technology is a subset of the achievement attribute which we wish to infer. Inherently, there is a problem of validity of inference unless the attribute of interest is very narrowly construed. MC tests then generate highly precise estimates of a portion of the attribute of interest and thus trade validity for reliability and efficiency.

Performance-based assessments, most of which require more obvious forms of human judgment in arriving at a score, are nonetheless typically more like the achievement attribute of interest (ability to think and reason; integrate skills and judge how to respond to tasks, etc.). A well-written performance item should have higher validity than a well-written MC item, or even several MC items, for no other reason than that the attributes it assesses are more nearly isomorphic with the constructs of interest. But to obtain similar levels of reliability (objectivity might be closer to what most folks are thinking about) for measuring individual status/performance, it takes several tasks (sampled much like MC testing but probably not as many items per topic/domain) and multiple judges (unlike MC) and thus can be more time-consuming and costly.

Implications for Assessment Design Given Purpose

The magnitude of the tradeoffs between MC and performance assessment depends on whether one is concerned with school/district/program accountability or student accountability. With respect to the former, the tradeoffs need not be as harsh between validity and reliability. By sampling multiple tasks and having multiple judges but not requiring all students to take all tasks and all judges to score all student responses, one can obtain high reliability at the group (school, program, teacher) level with reasonable student time demands and probably manageable costs for scoring. Programs like California Assessment Program and work by Bock and Mislevy (Bock & Mislevy, 1988; Bock & Zimowski, 1991) lay out how to do this.

With respect to student accountability, a considerable amount of adaptive assessment technology is needed to make the performance-based assessment process efficient and manageable unless it were to rely on the natural artifacts of

students' academic work (e.g., the interest in student portfolios as in Vermont [Vermont Department of Education, 1991]) or on "planted" standard tasks within ongoing instructional streams (part of the design scheme for the new CAP [California Assessment Policy Committee, 1991]). Even then, the design for collecting "artifacts" and developing a system for turning the heterogeneous mix into judgments will be complicated and costly, at least during the phase-in period. With respect to the adaptive aspects, it all boils down to quickly gauging student level of functioning and focusing the assessment tasks at that level. We have begun to see that occurring with MC testing, but the transfer to alternative assessment will be complex.

Implementation Issues in Alternative Assessments

There are problems to solve in implementing performance assessment per se. In using open-ended questions, there are design concerns about prompts (e.g., in mathematics this translates into how much guidance the prompt gives the student about what the desired response should entail; the new constructivist approach to mathematics argues for very little guidance because students should actively "construct" and choose their own meanings), design concerns about equivalency of questions (how does one make two essay questions equivalent? Does it depend on how many dimensions underlie the problem/question?), and concerns about scoring rubrics and scales (how much guidance in scoring? Should all questions be scored using a scale with the same number of dimensions or should the number of gradations vary according to how many dimensions underlie the problem?).

The new science performance tests where students conduct small experiments have a host of technical and practical problems in terms of their use in large-scale assessment. In California's 1990 pilot at Grade 6 wherein a hands-on science assessment of five tasks was group administered to classes of students, administration conditions and issues of group vs. individual testing were immediate concerns. Development costs are high for a set of tasks good enough and broad enough to avoid easy and early corruption. Scoring is very complicated as well.

Before we move too far to implement performance assessments on a broad scale, we had better systematically tackle some of these technical and feasibility problems. For example, the earlier mentioned Shavelson et al (1991) study of

science assessment would focus on one element, developing and demonstrating a strategy for creating alternative assessment tasks to tap the same key ideas without overly encouraging teaching to the task rather than to the key idea. Right now there are very few hands-on tasks at all, and if they were used both in instruction and accountability assessments, it is highly likely that teachers would prescriptively tune students to the assessment tasks themselves rather than to the concepts the tasks were designed to measure.

While it is straightforward to generate comparable tasks in the sense of classical parallelism (by using different ingredients in CAP's bags of circuitry for their electrical conductivity task or different material to test once the conductivity tester is constructed), these variants will likely be too close for us to feel comfortable about using the same task structure for both classroom and accountability purposes. Instead, we talk about creating "substitutable" assessment tasks. That is, tasks that measure the same substantive key ideas and concepts but have different task-specific attributes. Such tasks, given to randomly parallel sets of students, would likely have different performance distributions but the differences could be adjusted by statistical equating without unduly violating the principle of equitable measurement opportunities. We simply don't know how difficult this effort will be, but it is a necessary one given the multiple purposes performance assessments will be asked to serve in coming years.

Curriculum Change and Alternative Assessment

Curriculum forces and the broader policy community are committed to shifting from solely MC to assessments that on the face of it represent more important knowledge and skills closer to the attributes of interest. As long as assessment is going to drive instruction, why shouldn't better assessments (by which they mean more like the learning/knowledge/abilities of interest) do the driving? So the question is not whether we will have alternative assessment but under what conditions and when. The transition will be very rough.

A lot of people will have to learn a whole new way of thinking about assessment and many of them are classroom teachers and building administrators (Pandey, 1991). The staff development implications simply can't be ignored or we will have another "New math" mess (better curriculum but no one properly trained to use it). The question is how can accountability

assessment encourage this transition without being overwhelmed and unduly expensive in the process. The one inkling that progress can be made here is that the performance assessments can more readily engage teachers in the assessment process, either as external scorers of the new assessments or through embedding at least parts of the new assessments in regular classroom activities and using a student's teachers (through a moderation process) as primary judges. Over time, the staff development efforts to achieve higher assessment fidelity could indirectly benefit the teachers' routine assessment practices and thus tune them and, as a result, their students, away from a strictly MC mind set. Again, wishful thinking at this point, but perhaps a glimpse of the future.

Staged Change and A Mixed Portfolio

The way to move to performance assessments for accountability purposes is to tackle the necessary changes in stages. This new wave of assessments can't come about through immaculate conception. We are crossing a fairly deep chasm to reach a possible new horizon by the turn of the century (sooner, some hope). Accountability assessments that now rely strictly on MC tests will evolve, adding writing assessments, then "true" open-ended items in subject areas while gradually reducing the proportion of the test devoted to more traditional MC items (NAEP plans in reading and math for the 1992 cycle adopt this strategy; California's plans for the rest of the decade have a similar flavor). There will be a mixed portfolio of performance and MC tasks for a long time that may represent the final resting place in most cases. The commercial testing industry is already preparing for this eventuality as examples like Arizona and Maryland portend.

Portfolio assessment may not enter the school accountability arena in a big way right away but will likely be part of student accountability eventually. As pointed out above, most of the efforts to introduce performance-based assessments in state testing programs actually heavily involve teachers as rater/judges because of the benefits for fast staff development and spreading ownership.

Implications for Reporting Results

One hope is that the move to alternative assessment would encourage a movement away from trying to portray performance on a single score scale. Most performance-based exercises call upon the student to execute an array of content-

specific, metacognitive, and reasoning skills and to employ communication skills to boot. Given these features, adapting reporting strategies to treat tasks (MC or performance) as multiple signals of the individual's or the group's functioning would be desirable. While this might go against the grain of recent policy pressures to simplistic score reporting, surely it is possible to prepare the public to handle more complex but realistic messages. I certainly hope so.

Rethinking Notions of Validity of Educational Assessments

My CRESST colleagues and I are convinced that one consequence of the press for performance-based assessment for accountability purposes is a necessary rethinking of the measurement community's notions of validity of educational assessments. During the drafting of the CRESST assessment proposal, we decided that it was essential to spell out a broader set of criteria for judging the validity of educational assessment than had been used traditionally and conventionally. These criteria (proposed primarily by Eva Baker and Bob Linn within CRESST but motivated by ideas from Sam Messick, Lee Cronbach, and others), to be refined and modified through the course of the CRESST award, were intended to focus attention on the consequences and character of assessments as well as on more traditional technical and practical issues. As spelled out at the time of the proposal (see Linn, Baker, & Dunbar, 1991) for a more recent version), the criteria, beyond traditional concerns with reliability and validity for specific purposes, include attention to:

Assessment Consequences
Fairness
Transfer and Generalizability
Cognitive Complexity
Content Quality
Content Coverage
Meaningfulness
Cost and Efficiency

At CRESST, these criteria serve both as foci for research and as broadly applicable new standards which, we hope, others can use to evaluate their assessment alternatives. They underscore our recognition (belief) that we have to think differently about the validity of assessments in coming years and our

interest in improving the quality of current practice. Over time our thinking may change but at present the strong message is that notions of measurement, its purposes, its qualities, its consequences, and its uses are evolving, for better or for worse.

Concluding Comments

In trying to come to grips with the new assessment agenda and the complex policy and practice world in which it must unfold, I keep reminding myself how strongly I agree with a quote from the *Underachieving Curriculum* (McKnight et al., 1987):

> Complex enterprises generate complex problems requiring equally complex solutions. Schooling is such an enterprise. Therefore solutions to problems must, inevitably, be complex. . . . The longing for simplicity in the face of essential complexity is likely to produce deceptive explanations that lead to ineffective solutions. (p. 51)

Try substituting the words assessment, accountability, or whatever, in the quote. No matter what you choose in the measurement business these days, the message comes out the same.

The new world of assessment won't come about by hope or prayer, nor will the transition or its end result necessarily be tidy. But anyone who thinks it should be, as a feature of monitoring educational progress in an American-style democracy, is just dreaming of Lake Woebegone of the 1950s. Beaver Cleaver had better grow up; the new American pluralism and the increasingly global society have arrived on the educational measurement frontier.

Bibliography

Bock, R. D., & Mislevy, R. (1988). Comprehensive educational assessment for the states: The duplex design. *Educational Evaluation and Policy Analysis*, *10*, 89-105.

Bock, R. D., & Zimowski, M. (1991). *Individualized educational assessment: twelfth-grade science* (CSE Tech. Rep. No. 324). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

California Assessment Policy Committee. (1991). *A new student assessment system for California schools* (Executive Summary Report). Sacramento, CA: Office of the Superintendent of Instruction, California Assessment Policy Committee.

Learning Research and Development Center & National Center on Education and the Economy. (1990). *Setting a new standard: Toward an examination system for the United States*. Pittsburgh, PA: University of Pittsburgh, LRDC.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15-21.

McKnight, C. C., Crosswhite, F. J., Dossey, J. A., Kifer, E., Swafford, J. O., Travers, K. J., & Cooney, T. J. (1987). *The underachieving curriculum: Assessing U.S. school mathematics from an international perspective*. Champaign, IL: Stipes Publishing Co.

Pandey, T. (1991). *A sampler of mathematics assessment*. Sacramento, CA: California Department of Education.

Shavelson, R. J., Comfort, C., Barfield, D., Burstein, L., & Haertel, E. (1991). *Developing hands-on activities for assessing science understanding in diverse student populations,* A proposal to the Dwight D. Eisenhower Grants competition. [University of California, Santa Barbara]

Special Study Panel on Education Indicators. (1991). *Education counts: An indicator system to monitor the nation's educational health*, Washington, DC: National Center for Education Statistics.

U.S. Department of Education. (1991). *AMERICA 2000: An education strategy*. Washington, DC: Author.

U.S. Department of Labor. (1991). *What work requires of schools: A SCANS report for AMERICA 2000*. Washington, DC: U.S. Department of Labor, The Secretary's Commission on Achieving Necessary Skills

Vermont Department of Education. (1991). *Looking beyond "The Answer": Vermont's Mathematics Portfolio Assessment Program. Pilot year report 1990-91*. Montpelier, VT: Author.