CRESST REPORT 772

Eva L. Baker

WHAT PROBABLY WORKS IN ALTERNATIVE ASSESSMENT

JULY, 2010

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

Graduate School of Education & Information Studies
UCLA | University of California, Los Angeles

What Probably Works in Alternative Assessment

CRESST Report 772

Eva L. Baker
CRESST/University of California, Los Angeles

July, 2010

Reprinted from paper originally presented at the Annual Meeting of the
American Educational Research Association, Chicago, IL, April, 1991

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was conducted with partial support from the U.S. Department of Education, Office of Educational Research and Improvement, pursuant to Grant No. G0086-003.

The findings and opinions expressed herein are those of the author and do not necessarily reflect the positions or policies of U.S. Department of Education or the Office of Educational Research and Improvement.

To cite from this report, please use the following as your APA reference:
Baker, E. L. (2010). *What probably works in alternative assessment.* (CRESST Report 772). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Reprinted from paper originally presented at the Annual Meeting of the American Educational Research Association, Chicago, IL, April, 1991.

WHAT PROBABLY WORKS IN ALTERNATIVE ASSESSMENT [1, 2]

Eva L. Baker

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
University of California, Los Angeles

Abstract

This report provides an overview of what was known about alternative assessment at the time that the article was written in 1991. Topics include beliefs about assessment reform, overview of alternative assessment including research knowledge, evidence of assessment impact, and critical features of alternative assessment. The author notes that in the short term, alternative assessment will generate negative news about student learning and will require massive support to make it a successful reform strategy.

Introduction

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 educational achievement should reflect the complexity of learning which has created enormous opportunity to reform education through providing a focus on curriculum, staff development, and instructional improvement (Ambach, 1991; California Assessment Program [CAP], 1991; Baron, 1990; Resnick, 1990).

Beliefs

A prevailing, but as yet unsubstantiated, view of assessment reform has rapidly developed. This position holds that drastic changes in the nature of assessment—away from molecular, multiple-choice formats and towards more complex, meaningful, and integrative performance tasks—will result in concomitant improvements across the full range of educational activities. Success in reform through assessment implies a dramatic redeployment of expectations, resources, and the everyday events of teaching. These changes cannot be expected to occur naturally nor to be simply added on to the regular expectations of teaching. Such assessment will present intense challenges to standard views of the

[1] Originally presented as a symposium in *Authentic Assessment: The Rhetoric and the Reality*; conducted and augmented by a paper presented at the Annual Meeting of the American Educational Research Association, April, 1991, Chicago, IL. Data amended from version presented at the National Conference on Alternative Assessment; The Promise and Peril of Alternative Assessment, OERI, October, 1990, Washington DC.

[2] Thanks for Harold O'Neil, Jr., Yujing Ni, and David Niemi for their assistance.

curriculum, of ordinary teaching practices, and of the presentation of student achievement information to policymakers and to the public. But before such reforms can be made, or challenges faced, or impact assessed; high quality assessments must be systematically developed and validated.

Knowledge Base: Paltry But Sure to Improve

At the same time that interest in alternative assessment is high, our knowledge about the design, distribution, quality and impact of such efforts is low. This is a time of tingling metaphor, cottage industry, and existence proofs rather than carefully designed research and development. Urgency is in the air. Because many of these new assessments are being developed under constrained conditions of practice, for instance, by state departments of education or by local school districts, attention has focused on feasibility and schedule more than on technical quality of such measures. Moreover, because psychometric methods appropriate for dealing with such new measures are not readily available, nor even a matter of common agreement, no clear templates exist to guide the technical practices of alternative assessment developers (Linn, Baker, Dunbar, 1991).

Evidence from the Education Literature

In a systematic review of the literature in performance or alternative assessment, I reported on the amount of activity in research and professional literature (Baker, 1990). The database was first searched in June 1990 and resulted in 16,353 entries. By December, 1990, 17,475 entries in the ERIC system were found, an increase of about 7% in 6 months. We reviewed studies in the grade levels and subject maters reported in Table 1. The analysis indicated that more assessment studies were conducted at the secondary level than at the elementary level, in a ratio of 3:2. We also judged whether the studies focused on technical issues, for instance, validity or reliability. The distribution of effort by subject matter and by whether studies included empirical data is presented in Table 1.

Table 1.

Distribution of Performance or Alternative Assessment Studies Retrieved from ERIC (1983–1990)

Subject area	Number of entries	Empirical
Mathematics	304	15
Science	186	9
Social Studies	60	7
Writing	154	42
Foreign language	26	3
Second language	130	21

Given that the level of empirical work is so obviously low, one well might wonder what these studies are about. Some studies argue for new approaches to achievement testing. Others describe the development of a new measure for a particular type of student and/or for a particular subject matter. The most useful studies in the educational database were those conducted in the writing assessment area. Taken together, the educational studies result in the tentative truths shown in Table 2.

Table 2.

Generalizations from the Educational Literature

1. Raters can be trained to score open-ended responses reliably and validly.
2. Validity and reliability can be maintained through use of systematic procedures.
3. Training reduces the number of required ratings and costs of large-scale assessment
4. There is disagreement about the value and appropriateness of domain-specific scoring vs. more general scoring schemes.

Evidence from Military Literature

A second set of literature pertinent to the quality of the knowledge base for alternative assessment was retrieved from the National Technical Information Service (NTIS) database. This set focused on performance assessment studies from the military research and development. 14,774 entries were found. Of these, 187 dealt with raters (a key element in performance assessment), and only 41 provided empirical validity data. The findings from the military focus, in large measure, on the feasibility of alternative assessment for certification purposes (see Table 3).

Table 3.

Generalizations from the Military Performance Literature

1. Testing can occur in real or well-simulated contexts.
2. Large scale performance assessment can occur.
3. Assessments of both complex problem solving and team or group performance can be made.

As in the educational literature database, most studies from the military are descriptive and discuss how and why a different assessment technique is supposed to provide an advantage. Because the military has been conducting performance testing a longer time, the military database search yielded a rich set of assessment strategies. Table 4 lists some of these with their frequencies.

Table 4.

Assessment Formats from the Military Database

Assessment type	Frequency
Check-lists	450
Situational tests	27
Simulation tests	177
Hands-on tests	10
Skills tests	20
Performance samples	4
Unit, team or group performance	411

From my own analyses, the military experience provides some critical insights into how the alternative assessment movement might progress in public education. First, there has been little effort in the area of the validity of measures. Secondly, the provision of public or widely disseminated standards can corrupt the assessment process. My favorite personal experience is an isolated example where the complex task of tuning radio frequencies was reduced to a simple perceptual task (because the correct position for tuning was premarked by nail polish). The practical reason for such a change was clear. The tuning task was described as the hardest of the set of performance tasks, and a cue was necessary to assure that the appropriate numbers of trainees reached the criterion standards. A third observation inferred from military-based performance assessment experience is the tendency to reduce many problem-solving tasks to procedures. This transformation occurred to permit the performance to be judged simply as on or off in a checklist. Transforming problem-solving to procedures is certainly an appropriate strategy for technical tasks needing automatic responses. Yet, it is a curricular rather than a testing decision. Fourth, because it is organized hierarchically, the military assumes competence of raters who hold higher rank. Because of status, they are regarded as definitionally competent to make judgments. The result is a low investment in rater training and is a worry.

Evidence from Personal Research Experience

Evidence also comes from a series of studies (Baker & Clayton, 1989; Baker, Freeman & Clayton, 1991; Baker, 1990; Baker, Niemi, Gearhart, & Herman, 1990; Baker Aschbacher, Niemi, Chang, Weinstock, & Herl, 1991; and Baker & Niemi, 1991). These studies provide support for the assertions shown in Table 5.

Table 5.

Evidence from CRESST Content Assessment Studies (History and Science)

1. Using specifications, comparable tasks can be generated.
2. Expert-novice contrasts are a useful source of scoring criteria.
3. While tasks need contextualization, scoring rubrics can be formulated more generally, so that they are useful across topics and subject matters.
4. Teacher raters can be trained to be reliable, valid, and economical scorers using these rubrics.
5. Multi-step assessments can be implemented.
6. Construct validity evidence for the unique contribution of alternative assessment performance is emerging.

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 and the pressure to increase scores negatively influenced teaching and distorted curriculum (e.g., training in test-wiseness and a molecularized curriculum); testing can be a positive influence on the instructional behaviors of teachers. One commonly cited source of evidence for this assertion is performance in writing assessment. The reputed impact of the implementation of the California Assessment Program (CAP) writing assessment provides one such example. Data from the San Diego School District suggest that writing performance has dramatically improved on most types of writing assessed by CAP over the last 3 years (Raines & Behnke, 1991[3]). Yet assessment alone was probably not the only reason for such growth. As the Raines and Behnke report suggests, considerable efforts in staff development were made in parallel with 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 postsecondary colleges and universities to support improved instruction in writing, for example, through the California Writing Project. The conceptual and, to some extent, procedural analyses requisite for the design of staff development preceded the CAP writing 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?

[3] Thanks to Grant Behnke and the San Diego City Schools for making this research report rapidly available to me.

Issues and Predictions

Despite this fragile research base, alternative assessment has already taken off. What issues can we anticipate being raised by relevant communities about the value of these efforts? What problems are ripe for research? Where are we now?

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 the assessment task has more implicit meaning to students. This interpretation is rich in research opportunities. Alternative assessment means anything but multiple-choice (and true–false) problems but generally connotes extended and multistep production tasks. Such tasks inevitably require the use of raters, judges, or their electronic proxies to determine the quality of students' efforts. 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 become more precise. They must include the designation of the type of intellectual skill assessed (such as explanation or problem-solving). Appealing format changes do not assure serious attention to higher-level skills. A portfolio is not a portfolio is not a portfolio. We need to hurry the process through which 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; and second, how will the assessment be used? To the first point, if an assessment is to serve in any way as a standard for individuals to demonstrate competency or to provide a mark for system performance, desired intellectual processes and content/situation domains must be clearly and explicitly 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? Although general frameworks such as the National Council of Teachers of Mathematics (NCTM) standards provide a point of departure, they are far from precise enough neither to design appropriate measure nor to create targeted instruction. In any case, some explication of the intention and class of performance, which the assessment task represents, must be described. This structure assumes that at least alternative assessments attempt to provide a general framework in which to understand students' accomplishments, even in the absence of agreed-upon standards. Task specification seems an obvious option (Baker, Niemi, Aschbacher, Ni, & Yamaguchi, 1991).

The second issue, assessment purpose, forces the consideration of questions of representativeness of student performance on alternative assessments. Given the extended time periods and resources required by many alternative assessments, we need to feel that our findings are trustworthy and fairly represent student capability. Recent research (Shavelson, 1990; Linn, 1991; Baker et al., 1991), and pronouncements (Hoover, 1991), suggest that task sampling is a major validity issue. Specifically, researches have found only moderate correlations for individual students' performance across different tasks in the same subject matters. 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? Until some insight on this phenomenon can be developed, however, using a single performance assessment for individual student decisions is a scary prospect. We are unlikely to be able to trust our results.

Format and Criteria: Two Critical Features of Alternative Assessment

I have noticed a disconcerting tendency to overvalue differences in format, (e.g., hands-on, portfolio, multistep 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 to assemble groups of teachers to decide on scoring dimensions. Evidence from our own research, however, suggests that teachers are not necessarily 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 the performance of experts in contrast to those of 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 scoring criteria include students' use of prior knowledge, principles, newly acquired information, and avoidance of misconceptions; 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 of assessments developed with different models.

Models for developing scoring criteria, however, involve more than technical concerns. Within some fields, such as writing or history, there are ideological differences of opinion regarding which set of criteria are appropriate 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 has proposed to conduct 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, common understandings and common standards for performance for both accountability and instructional purposes are required if equity is to be served and performance disparities reduced. If students in different schools are held to vastly different types of performance, equity issues will exponentially increase with performance assessment.

Alternative Assessment is Going to Generate Bad News in the Short Run

Our research in developing alternative assessments in history and science shows that 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 do not know how to do what is expected of them in these tasks. The dilemma is that we cannot improve the quality of these assessment 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 do not 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 administration, and so forth. I believe that people do what they know how to do. And I imagine that many teachers simply do not 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 and resources are scarce. 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 download staff development responsibilities, including the continuing design and scoring of assessments, to local districts. We may have even a bigger problem, for redesigned staff development assumes we know what we want to teach teachers to do and how to teach them effectively.

Beyond resources for assessment and staff training, systematic development, and implementation of alternative assessment have additional costs. On the mundane level, teachers have told us they need additional teaching assistant time simply to manage students during alternative assessments themselves, let alone to 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. Secondly, 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. Thirdly, 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 their homes. 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 kids 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 and more persistent meaning than the narrowing of multiple-choice performance.

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 as 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. At best, the lack of student interest may be a transitional problem, ameliorated following exposure to appropriate instruction.

Assess Smarter

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 make no sense without thinking in parallel about real problems—about issues like 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 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.

Validity Criteria for Alternative Assessment

Just the technical assessment problems alone can occupy many of us until the next century. But we clearly must emerge from narrow specialization if we are going to grapple with the broad implications of alternative assessment reform. Clearly the technical agenda for alternative assessment requires attention. Of particular interest is the extent to which alternative assessments developed in different locations provide comparable information about student accomplishments. A second major concern is quality. How do we judge the quality of alternative assessment. We have proposed (Linn, Baker, & Dunbar, 1991) an expanded set of criteria to use to evaluate the validity of new assessments. These criteria focus on two major categories: properties of the assessment itself and factors external to the assessment.

Internal Validity Criteria

The first internal criterion by which alternative assessments should be judged is the **Cognitive Complexity** that assessment demanded by the measure. This criterion requires looking beyond surface features, such as hands-on experiments or paper and pencil problem-solving. Rather, one must attend to the intellectual demands of the task. The determination of cognitive complexity also requires knowledge about students' prior instructional experiences, to assure that a task that appears to be novel had not in fact been practiced and memorized sometime before the assessment period. A second criterion is the **Meaningfulness** of the task. This criterion encompasses concerns that the directions for the task are clear and expressed in language or symbols generally accessible to students. More importantly, however, the meaningfulness dimension addresses the extent to which the task is contextualized or framed in a setting or in specific examples which students are likely to understand. From this second perspective, meaningfulness relates both to attention and motivation to perform and to issues of fairness and bias. Third, any estimate of alternative assessment quality should attend to the content **Quality** inherent in the assessment. Subject matter content must be expressed accurately and to prevailing standards in the discipline. A knotty component issue to content quality is the extent to which the assessment represents a range of different topics (breadth) or focuses more narrowly on a limited set of topics. The importance of this dimension grows if the alternative assessment is intended to be used for making individual decisions about students, for then the issue is raised of the limits of task sampling, or the representativeness of the assessment content with regard to the domain. A fourth criterion focused on the assessment itself is the extent to which the assessment provides for, enhances, or encourages the generalizability and transfer for the particular performance to other related topics or domains. **Transfer** and **Generalizability** are enhanced by tasks that are conceived to be relevant to a wide range of applications and by the use of criteria for judging performance which themselves have wide applicability across tasks. Generalizability would be encouraged by the public articulation of standards for both task generation and generalizability so that performance criteria could be learned and articulated by the student.

External validity Criteria

Three external validity criteria should be used to assess the quality of alternative assessment efforts. The first criterion in this category is **Consequences**, a term derived from the writing of Messick (1989). As we interpret this criterion, consequences encompasses two major concerns. First that the consequences of the assessment for the student are appropriate—people who deserve to pass, do so. We would also expect to see regular patterns of performance among range of achievement and other measures of student performance. A second, more difficult interpretation focuses on the impact of the assessment on a wide range of other educational elements, a concept explored by Frederiksen and Collins (1989). Do the assessment findings lead to desired ends? For instance, do high standards result in higher performance for all students or disproportionate dropout rates? Does it result in improved and challenging curricula or assessment-focused practice exercises; more motivated and interested students or less attentive participants; or stimulated and renewed teachers or stressed and demoralized teachers? Certainly no assessment should reap the reward or bear the blame for all of these options. And conducting definitive studies of impact may present a fascinating technical challenge. However, the intent is to attend to a broader range of policy consequences when implementing major assessment changes so that any unanticipated negative outcomes can be identified early and remedied expeditiously.

Closely related to the consequences criterion is the criterion of **Fairness**. We want to assure that the assessment is fair in a number of ways. First, we need to assure that students will not be disadvantaged because of characteristics irrelevant to those of interest in the assessment. Performance should be independent of the student’s gender, race, social class, ethnicity, or site of residence. We should also assure that we can document that students have had a reasonable opportunity to learn desired material to assure some level of adequate preparation. Finally, we need to be sure that scoring procedures and raters are free of bias or differential expectations, a problem of concern earlier noted. Finally, alternative assessments should be judged in terms of cost and efficiency. We know that such assessments require more resources in time to develop, to administer, and to score. We need to assure ourselves that the quality of performance measured by such assessments is more credible and desirable, and that the assessment provides multiple kinds of information about the student to warrant the cost. Clearly, if alternative assessments have the multiple benefits suggested by their proponents, then they will be worth most any price.

Conclusion

No one could have predicted how rapidly two parts of the educational world would converge: the exploration of new assessments and the growing consensus that a major national revision of assessment is critical to our future. This fast-track interaction of technical possibility and policy imperative presents enormous challenges to the technical community, to the concept of assessment, and to goodwill among policymakers, technical experts, schoolteachers, and administrators. Mike Rose (1989), a UCLA colleague, wrote a book about underprepared university students. His quote seems appropriate to our times.

> We are a nation obsessed with evaluating children, with calibrating their exact distance from some ideal benchmark. In the name of excellence we test and measure them…and we rejoice or despair over the results. The sad thing is that though we strain to see, we miss so much…those most harshly affected…possess some of our greatest unperceived riches.
>
> Mike Rose, *Lives on the Boundary*, 1989

Let’s work together to attempt to perceive some of these hidden assets.

References

Ambach, G. (1991 March). *Improving curriculum and instruction*. Paper presented at the UCLA/CRESST conference Educational Assessment for the 21st Century: The National Agenda, Los Angeles, CA.

Baker, E. L. (1990, April). *Assessment and public policy: Does validity matter?* Paper presented at the American Evaluation Association Annual Meeting, Boston, MA.

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

Baker, E. L., & Clayton, S. (1989). *The relationship of anxiety and performance in content-based higher order thinking.* Paper presented at the 10th Annual Conference of the Society for Test Anxiety Research, Amsterdam, The Netherlands.

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., & 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 National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., & Niemi, D. (1991, April). *Assessing deep understanding of science and history through hypertext*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

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., Niemi, D., Gearhart, M., & Herman, J. L. (1990, April). *Validating a hypermedia measure of knowledge representation*. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA.

Baker, E. L., & Stites, R. (1991). Trends in testing in the USA. *Politics of Education Association yearbook*. London: Taylor & Francis.

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

Burstein, L. (1991, April). *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, IL.

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

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

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

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

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd. ed., pp. 13–103). New York, NY: Macmillan.

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

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, DC.

Rose, M. (1989). *Lives on the boundary*. New York: Penguin.

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

Torney-Purta, J. V. (1990, April). *Measurement of Performance in Social studies*. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Wolf, D. P. (1990). *Assessment as an episode of learning: Making new assessments broad enough*. Presentation to the Department of Education's Conference on Alternative Assessment.