Research Report

HIGH-STAKES & assessment INNOVATION

a NEGATIVE correlation?

Sri Ananda and Stanley Rabinowitz

WestEd
Improving education through research, development, and service

WestEd
Improving education through research, development, and service

Research Report

HIGH-STAKES & assessment INNOVATION a NEGATIVE correlation?

Sri Ananda and Stanley Rabinowitz

No period in history was more eventful for assessment than the last decade of the 20th century, when several developments seemed to signal a revolutionary change in how student learning and achievement would be measured. Take the case of performance-based assessment, which seemed destined for a perfect marriage with standards-based reform. With rigorous, new academic standards identifying what students were expected to know and be able to do came the need for innovative assessment methodologies that could gauge student knowledge and performance with greater richness and depth. Performance-based assessment seemed like an ideal innovation to support standards-based reform.

But what had seemed in the early 1990s like a model relationship no longer looked so promising to state-level policymakers as they headed into the 21st century. What came between standards-based reform and innovative assessment was the proliferation of statewide accountability systems that rely heavily on large-scale testing to measure student achievement, and on various rewards and sanctions to motivate educators and learners alike. The attachment of high stakes to test results fueled increased scrutiny of the tests themselves: Were the assessments valid, reliable, and fair enough to carry the weight of accountability? Were they affordable in a large-scale setting? Did they allow for timely dissemination of results to key stakeholders?

Today, several states have scaled back or delayed envelope-pushing assessment formats and systems. In fact, contrary to predictions made a decade ago by some proponents of performance-based assessment, standardized norm-referenced testing continues as the reigning methodology of large-scale assessment.

This paper argues that, as implemented thus far, there has been an inverse correlation between innovation and accountability in statewide assessment systems: the higher the stakes attached to assessment results, the more conservative the assessment methodology ultimately used. Included below are case studies of two state assessment programs that illustrate 1) the increasing and often overwhelming demands for accountability

throughout the education system and 2) the inadequacy of existing assessment delivery infrastructure and methodology to easily accommodate innovation. The paper concludes with a series of "lessons learned" that offer hope for our ability to develop and implement more effective, efficient assessment systems, even as we rely more heavily on them for accountability purposes. These lessons reveal that, when carefully conceived and implemented, innovation and high stakes *need not be* mutually exclusive. This message is particularly timely as a number of states begin to explore still newer assessment innovations, such as computer-based and on-line testing, for their assessment-and-accountability systems.

The Stumbling Blocks for Innovation

Almost counter-intuitively, the proliferation of state accountability systems that attach high stakes to test outcomes has worked against the use of innovative assessment at the state level. As states have moved to implement "world-class" standards for their students, the assessment systems used to measure progress toward achievement of those standards have changed less than expected and less than many have advocated. Few would have predicted this persistence of the status quo because the intent and substance of many newly developed state content standards seemed to cry out for assessment innovation. Proponents of performance assessment argued that traditional statewide tests — consisting primarily of multiple-choice items, sometimes accompanied by a limited number of short-answer items — couldn't adequately capture the rich learning implied in the standards. The traditional formats were regarded as limited in their ability both to measure complex, higher-order content and to drive improvements in teaching and learning. Yet, despite this clear rationale and the resulting enthusiasm for ramping up performance assessment in statewide systems, the effort stalled with the introduction of high-stakes accountability. Although a number of states had begun planning and implementing innovative[1] assessment models, many of them performance-based, the introduction of high stakes — combined with some faulty premises and unrealized promises related to performance assessment — provoked a conservative backslide. Nagging concerns undermined even the most determined efforts. These concerns fell into several interrelated categories: technical limitations, logistics, professional development, cost, and political will. Each will be discussed below.

Technical Limitation. Contrary to early expectations, compared to traditional testing formats (e.g., multiple choice), many innovative assessment formats were found to have lower reliability and generalizability indices as commonly calculated. In response, proponents of innovative assessment called for new ways of defining reliability and validity. But, given the rewards and sanctions associated with new state

[1] While no single definition of innovative assessment is preeminent, states have commonly referred to direct writing assessment and other constructed-response item types as examples of innovation in their programs. While such item types represent an advance beyond multiple-choice assessments, they do not represent the vision of the adherents of performance-based assessment (Mitchell, 1992; Wiggins, 1993); such models extolled the virtue of curriculum-embedded assessments such as projects or portfolios. Others, such as California Learning Assessment System (CLAS), called for a greater reliance on "enhanced multiple-choice" formats — items that go beyond recall and measure higher-order skills to include problem solving and applications of knowledge across a range of situations. More recently, computer-based testing is emerging as a popular assessment innovation.

accountability systems, policymakers were wary of ignoring the old technical standards, particularly when some ambitious state systems began receiving negative external reviews by technical experts (e.g., Koretz & Barron, 1998). Further undermining enthusiasm for the innovative assessments was evidence that the relative under-performance by females and minorities on traditional tests persisted with the new assessment methods. Worse yet, in some cases, achievement gaps actually increased.

Logistics. One indisputable strength of standardized multiple-choice assessments is the relative ease with which they can be administered and scored and their results can be reported. Only moderate effort is needed to prepare students and teachers on the nuts and bolts of the testing process; and once that process is understood, the procedures generalize across grades and content areas. Innovative assessment proved harder to manage overall. As assessment moves from being a discrete event (e.g., taking a 50-item multiple-choice test) to being a "continuous" process (e.g., developing a portfolio over the course of a school year), increased planning and support structures for assessment implementation and scoring become essential. Many states were unable to develop the necessary infrastructure to support innovation (e.g., statewide student tracking systems, cadres of trained scorers, teacher support networks). In turn, inadequate infrastructure often contributed to logistical snafus and delays in assessment administration, analysis, and/or reporting of results. In other words, the trains didn't run on time, and when that happens in the high-stakes assessment arena, public confidence and support quickly dissipate.

Professional Development. Often acknowledged, but seldom adequately addressed, is the need for extensive professional development for teachers and administrators to support innovative assessment. For example, use of performance-based assessment requires significant changes in classroom structures and procedures, with teachers assuming more of a mentoring than a lecturing role. Performance-based assessment also requires that they translate standards into exemplar tasks and classify student work into categories of achievement (e.g., Basic, Proficient, Advanced). Yet the degree of resistance to such change is often underestimated, as is the amount of time and effort needed to adequately and positively support the required transformation of attitudes and skills. The challenge for all involved is exemplified in a letter written by a Kentucky teacher asking how was she was supposed to teach all of her content and have students do portfolios, too. Clearly, this teacher lacked the fundamental understanding that a portfolio is supposed to be a vehicle for teaching important content.

Cost. Innovation is expensive; and shifting from multiple-choice to performance-based formats is especially so. Costs increase in several ways. Developing open-response items can be as much as 10 times more expensive than developing multiple-choice items. Similarly, performance tasks often require more iterations of field testing than multiple-choice items. A constructed-response item can typically assess multiple standards more readily than can a multiple-choice item; yet reaching adequate levels of reliability and generalizability usually requires the inclusion of more constructed-response items than states have been able to afford financially. The cost of time must also be considered. For example, adding more constructed-response items would increase testing times, possibly beyond the point schools or students would tolerate.

Political Will. Faced with the various implementation challenges outlined above, many state policymakers have chosen to back away from assessment innovation rather than delay their accountability systems until their

education system could catch up to the new assessment models. Simply put, most states have opted for "results now." This collapse of political will is not surprising given the attention focused on some of the early adopter states that found themselves having to backtrack, delay, or, in a few highly publicized cases (e.g., Arizona, California), drop the new state assessment program in its entirety.

Casualties of Reform: Case Studies

As states have begun to backpedal on hoped-for innovation in their assessment programs, some have couched the changes as delay. In other instances, states have outrightly eliminated an innovation — entirely and permanently. Among the retreats have been:

- dropping innovative assessment formats (e.g., performance events in Kentucky and multiple-choice items with multiple correct answers in Pennsylvania);
- removing assessment items that link to higher-level or "world-class" standards (e.g., Arizona, California); and
- delaying implementation of assessments in challenging academic (e.g., science) or non-core academic content areas (e.g., workplace readiness).

Presented below are the stories of two states in which promising, innovative statewide assessment formats were rolled back and, ultimately, eliminated. We have selected these states because, like many others, they based their reform efforts on high standards for all students in both traditional and nontraditional content areas. Each effort fell victim to the scrutiny that naturally attends high-stakes systems and to a lack of readiness throughout the broader education system.

Case Study #1: Kentucky Instructional Results Information System

The Kentucky Education Reform Act (KERA, 1990) served as a wake-up call that the status quo would no longer be acceptable in the Commonwealth's public schools. At its core, KERA declared that all students could and must learn at high levels. As an incentive to meet this goal, educators would be rewarded or sanctioned, depending on students' achievement across a wide range of "Valued Outcomes" (later redesignated as "Academic Expectations") consisting of academic and noncognitive indicators.

The linchpin of KERA was the Kentucky Instructional Results Information System (KIRIS), a truly innovative, multi-method assessment program. KIRIS was designed to change classroom behavior, in recognition that the ambitious goals of KERA could not be reached unless fundamental reform in curriculum, instruction, and assessment infiltrated the entire education bureaucracy. The most important shift needed to occur in the relationship between students and teachers.

As originally conceptualized, developed, and implemented, KIRIS contained the following innovative features and assessment methods:

Open-Response Items. While students were administered both multiple-choice and open-response formats, reporting was based only on the open-response items. (The multiple-choice items were included for equating and other technical purposes.) This approach was intended to send the message that the type of instruction needed to prepare students for these more demanding tasks had to be at the forefront of all change; schools could not meet their accountability goals without emphasizing writing and problem solving.

Portfolios. In elementary, middle, and high school, students were expected to complete a writing portfolio as part of their classroom activity. Over time, a mathematics portfolio would be added at these same three levels.

Performance Events. A set of performance tasks was developed to assess students' ability to work in groups, problem solve, and summarize their findings. Trained facilitators were sent to schools to oversee student performance — to ensure comparable administration and test security.

Integrated Assessments. Items were developed to assess multiple content areas (e.g., mathematics and science; social studies and practical living) as a way to develop more complex, "naturalistic" tasks.

Noncognitive Indicators. In addition to the testing accountability components, schools would be rated on their achievement of nonacademic indicators including: attendance, retention, dropout rates, graduating students' "successful transition to adult life," and "reduction of physical and mental health barriers to learning."

KIRIS to CATS: Conservative Cutbacks

After a decade of reform, KIRIS was transformed to CATS (Commonwealth Accountability Testing System). While CATS retains some of the innovations that defined its predecessor, several significant changes occurred in its transformation from KIRIS, almost all of them moving away from innovation and toward the inclusion of more traditional components. Below, we briefly summarize these changes and explain the shift away from innovation.

Open-Response Items. Results from the multiple-choice items were added to reports and informed accountability decisions. This addition reflected three concerns: (1) making all items on which students are assessed "count," (2) criticism in some quarters about the scoring reliability of the open-response items; and (3) a need to create more variance at the lower end of the performance scale in order to more accurately gauge student achievement at this end of the scale. In addition, some of the original "Valued Outcomes" (e.g., student attitudes) were excluded from new testing because of questions about their appropriateness for a statewide testing program.

Portfolios. While the writing portfolio continues to be administered, plans for the mathematics portfolio were dropped for a combination of reasons: (1) concerns over the cost of administration and scoring; (2) lack of mathematics teachers' readiness to use instruction appropriate for a portfolio, especially developing tasks for inclusion; and (3) concerns about creating a burden on teachers and students.

Performance Events. Performance events turned out to be popular for many teachers because they supported cooperative learning and a problem-solving approach to instruction. Unfortunately, they were dropped from the assessment system for several reasons: (1) despite the presence of the facilitators, administration varied so greatly across the state that inclusion of these tasks lowered the reliability of the overall school accountability index; (2) performance events scores were relatively uncorrelated with those of other methods within content areas, making interpretation of proficiency difficult; (3) the logistics (training, scheduling, and delivery) of these tasks were so monumental that occasional problems (e.g., missing materials, absent facilitator) were inevitable, making it difficult to determine the appropriate adjustments to high-stakes school accountability scores; and

(4) the cost of performance events was difficult to justify given the many problems.

Integrated Assessments. Such tasks are no longer part of the assessment plan because: (1) not all content naturally integrated across subject areas — some tasks proved to be less authentic than desired; and (2) due to time and cost limitations, open-response tasks could not be scored separately for both content areas — inevitably, the "second" area was scored less reliably.

Noncognitive Indicators. "Reduction of physical and mental health barriers to learning" was never added to the accountability formula due to an inability to develop an overall definition that would apply across the Commonwealth. "Successful transition to adult life," while still a part of the accountability index, is limited to the first six months following high school graduation due to a variety of factors; this limitation has resulted in a much narrower definition of transition than originally envisioned.

Norm-Referenced Standardized Test. Accountability decisions now include students' reading, language arts, and mathematics scores on the Comprehensive Tests of Basic Skills at three points: exiting primary, grade 6, and grade 9. Inclusion of a national norm-referenced test, a growing phenomenon in many states, is intended to satisfy two concerns: (1) how do students perform relative to national standards? and (2) can the public believe the results of state-developed tests?

In conclusion, the high-stakes consequences dictated by KERA placed a burden on KIRIS that it could not carry. The major casualties of the mounting criticism were the more innovative assessment formats. These tasks could not survive several setbacks, including: (1) teacher complaints about burden and unfairness; (2) external technical reviews questioning their technical adequacy; and (3) logistical shortcomings.

Case Study #2: The Development of a Career-Technical Assessment System

As previously mentioned, venturing beyond the traditional core curriculum to include nontraditional content (e.g., workplace readiness skills) is among the more recent innovations in statewide student assessment systems. California's Assessments in Career Education (ACE) program is an example. ACE is a standards-driven, career-technical (vocational) assessment program for high school students, which was incorporated into California's operational statewide student assessment system in the late 1990s. Although it's not part of a formal statewide accountability system, ACE is considered high stakes for its target population because students who perform well on it receive recognition on their high school diploma. That recognition is valued by prospective employers, as well as by several postsecondary education programs.

However, ACE's incorporation has not been a smooth process; more than eight years passed from its initial development to its administration statewide. As with the Kentucky example, the existing ACE program is narrower in scope and more traditional in methodology than originally planned.

Beginning as the Career-Technical Assessment Program.

Two distinct movements in the early 1990s provided the context for the development of a comprehensive statewide career-technical assessment program: the movement to reform vocational education and employment training programs, and the emergence of performance-based education assessment techniques as

alternatives to multiple-choice testing. In 1990, work began on Career-Technical Assessment Program (C-TAP), with a contract from the California Department of Education (CDE) to WestEd. The initial emphasis was on the development of a series of occupation-specific multiple-choice tests. But within the first year of planning, C-TAP was completely reconceptualized as a standards-driven, performance-based student assessment system. This was consistent with both the vocational reform and alternative assessment movements. Its primary purpose was to certify and formally recognize students demonstrating mastery of important career-technical and academic competencies consistent with California's Model Curriculum Standards for programs in Agriculture, Business, Health Careers, Home Economics, and Industrial and Technology Education.

Given the history of hands-on assessment in vocational education, as well as the new emphasis on integrated and higher-order learning in education in general, C-TAP was seen as an ideal laboratory for investigating different types of performance-based assessment tasks targeted to challenging standards and higher levels of cognition. After developing and pilot-testing several different types of performance-based assessment tasks, the C-TAP model settled on the following combination of cumulative and on-demand components: a portfolio, a project (including product and oral presentation), and written scenarios (complex problems or situations presented in a career context to which students must propose a solution in writing). These three assessment methodologies were selected in large part because pilot-testing demonstrated that, compared to other performance tasks, they were the most likely to be effectively implemented in different school and classroom settings across the state. For purposes of certification, students were expected to complete all three assessment components.

Collectively, the C-TAP components were intended to provide different types of evidence of student learning relative to: (1) general workplace standards (Career Preparation Standards), (2) career area standards (Model Curriculum Standards), and (3) related academic standards. Furthermore, C-TAP was designed to be consistent with the direction taken by California's academic student assessment system under development at the time — the California Learning Assessment System (CLAS). CLAS was an ambitious, performance-based student assessment system aimed at satisfying many needs not met by the previous academic student testing program. Among them were assessment and reporting of individual student performance; alignment of assessment to content taught in schools; and more direct and meaningful measurement of content through performance-based assessment (Kirst & Mazzeo, 1996). Both C-TAP and CLAS had cumulative assessment components, including portfolios, as well as on-demand components. A noteworthy difference is that CLAS featured multiple-choice items, whereas the C-TAP model completely dropped the multiple-choice items at a very early stage of development.

The Demise of CLAS and Rethinking of C-TAP.

By Summer 1992, the C-TAP model was fully developed and had been pilot-tested in classrooms throughout California. However, by Fall 1992, the plan to expand the C-TAP model to other career areas was scaled back substantially. It was becoming increasingly clear that the level of resources necessary to support a statewide, performance-based, vocational student assessment system with components tailored to 20-plus career cluster areas had been significantly underestimated. After two years of intensive development effort, assessment materials were available for fewer than half the targeted career areas, and none of the

assessments was yet ready for statewide implementation. Moreover, information from field testing and other data collection efforts (i.e., teacher and student interviews and surveys) suggested that the C-TAP performance-based assessment system was perceived as burdensome for many individual teachers. For example, a substantial number of vocational teachers felt that they could not provide the writing instruction that many of their students needed to develop a portfolio (ETI, 1997). Finally, there was a lack of willingness to expend the political capital necessary to push ahead this ambitious assessment agenda.

As C-TAP began facing increasing obstacles to statewide implementation, CLAS was administered for the first time in Spring 1993. Controversy quickly followed. Parents and conservative groups expressed concerns about CLAS's "objectionable" content (e.g., some charged that the content invaded student privacy and others took issue with the controversial subjects touched on by some assessment items). These concerns were heightened by the California Department of Education's maintenance of test security. While that security was intended to protect the integrity of the test and to avoid the expenditure of human and financial resources for new development, critics perceived the standard security measures as a deliberate attempt to keep the public in the dark about test content.

In addition to the controversy over content, the assessment's sampling procedures came under criticism. Lawsuits were filed. The final blow seemed to come from the results of the commissioned evaluations, some of which were undeniably critical of technical aspects of the assessment system. California's governor ultimately called for the development of a new statewide assessment system (Kirst & Mazzeo, 1996). This blowout over CLAS was to have serious effects on the future of C-TAP.

C-TAP as a Model for Local Adaptation and Implementation.

Although C-TAP assessment development and field testing continued through 1993–94, the demise of CLAS in 1995 contributed to a formal change in program objective. The CDE decided that statewide implementation of C-TAP was politically untenable due to lack of public support for portfolio assessment, the high cost of administering and scoring, and insufficient evidence of the system's technical adequacy. But, at the same time, CDE acknowledged, the C-TAP model was popular with teachers and schools in pockets across the state. Thus, CDE decided to shift the overall purpose of C-TAP from providing an assessment *system* to support statewide student certification to providing an assessment *model* for local adaptation and implementation.

New state legislation also led to another significant change in C-TAP at this time. Assembly Bill 198 mandated that California students be "prepared to enter the work force." Many interpreted this to mean that students be required to learn general workplace readiness skills (e.g., teamwork, use of technology, use of information). Thus, the C-TAP model was expanded to incorporate generic workplace readiness assessment components, in addition to those components tailored for particular career areas. The new component, the Career Preparation Assessment (CPA), was aimed at both vocational and non-vocational education students. It was designed to feed into the C-TAP system so students could begin with the more generic CPA and, over time, build the specialized career-related skills needed to meet the more specific C-TAP requirements in their career area of interest. Alternatively, for students not enrolled in vocational or career-related programs, the CPA model could provide culminating evidence of their proficiency on general career preparation skills.

Besides new state legislation, the passage of the School-to-Work Opportunities Act of 1994 also helped to keep C-TAP alive in the post-CLAS era. The Act, which provided states and local school districts with "venture capital" to develop comprehensive school-to-work transition systems, called for portable skill standards and certification for students. The career vocational education division of the CDE thought that C-TAP could provide an assessment model for school-to-work skill standards certification.

Restructuring C-TAP as ACE.

In 1995–96, CDE faced a period of reorganization related to budget cuts and a resulting need to downsize. As part of that reorganization, C-TAP was moved from the career vocational education division to the assessment (i.e., student performance) division. This administrative move resulted in the most comprehensive changes on this assessment system to date.

The first major change made to C-TAP under the assessment division was to begin development of on-demand tests that comprised multiple-choice items and constructed-response tasks. The development and implementation of performance-based assessments, other than selected constructed-response tasks, was put on hold indefinitely as a result of the demise of CLAS. In addition, CDE's assessment division aligned the new career-technical assessment effort to its closest operational academic student assessment counterpart, the Golden State Examinations (GSE). Established in 1983, GSE offers end-of-course examinations in key academic subject areas to students in grades 7 through 12, and provides recognition to students who demonstrate outstanding levels of achievement on each examination. Thus, while the C-TAP portfolio, project, and written scenario components were made available for adaptation at the local level, efforts at the state level were redirected to development of a new career-technical assessment system, *Assessments in Career Education (ACE),* which more closely paralleled CSE in format and purpose.

Formal incorporation of ACE into the operational statewide student assessment system and adaptation to the GSE model required some accommodations to existing guidelines that aren't typically applied to career-technical assessments. For example, the political outcry over CLAS led to a policy decision that statewide assessments could not ask about or mention anything to do with personal or family beliefs or ethics; such questions were considered too personal and, thus, invasive. This policy presented a problem for the ACE in Health Careers: It is generally accepted in the health-care field — and codified in the national standards for health-care workers — that *all* workers in the field must be knowledgeable about ethical expectations and practices (e.g., patient confidentiality, patients' right to know). Thus, many ACE Health Care items were developed to assess proficiency with respect to this important standard. When the health-care test items were reviewed by CDE's legal specialists prior to placement on operational test forms, all items mentioning ethics of health-care workers were rejected because the reviewers were concerned about the requirement to avoid anything touching on personal or family ethics.

The ACE examinations (in Agricultural Core, Computer Science and Information Systems, and Health Careers) became operational for the first time in Spring 1998. They were administered to fewer than 10,000 students, a low number resulting from several factors: (1) the limited pool of students eligible to take ACE examinations (i.e., students enrolled in selected career-technical programs); (2) lack of concerted

public relations effort to inform the key constituents in the field (i.e., schools, teachers, students, parents, employers) about the purpose, scope, and benefits of these examinations; (3) confusion about which students were eligible to take the examinations; and (4) inadequate (less than two months) notice to the field concerning the window of test administration. Fortunately, the latter three factors were effectively dealt with in subsequent years of ACE administration.

Several years later, the ACE program remains a part of the California student assessment system. But given its history and current status, it is fair to say that the future of ACE seems uncertain. Even with the best of intentions, a state department of education would be hard-pressed to continue supporting an assessment program with such low student participation. However, despite the rocky road to incorporating career-technical education into its statewide student assessment system, California remains one in a minority of states to have actually achieved this.

Implications for Other High-Stakes Statewide Student Assessment Programs

How can these two case studies be helpful to states contemplating new innovations and higher stakes for their student assessment systems? They yield five major lessons:

Lesson #1: High-stakes assessment systems that are primarily performance-based may not yet be viable at the state level. The Kentucky and California experiences are not unique with respect to this issue. Whether they have dived whole-heartedly into the movement or merely "tested the waters," those states that have used this innovative methodology can attest to the resources and political will required to support performance-based assessment. Kentucky's performance events and math portfolios are no longer part of the statewide student assessment system because of technical considerations (e.g., decreased reliability) and logistical difficulties (e.g., insufficient resources for the extensive teacher professional development required to support the innovative assessments). Similarly, California's decision to relegate much of the responsibility for career-technical performance-based assessment to the local level was a practical response. But the "old" C-TAP portfolio model survives, and even thrives, in some form at many individual schools across the state.

Lesson #2: If there is widespread support for a particular assessment innovation, states will invest. In California, funds to support the early vision for statewide use of the portfolio-based career-technical assessment system never materialized, largely because it was not a major assessment priority for the state. However, when there is strong support for an innovation, states have demonstrated a willingness to invest. For example, 20 years ago, open-ended writing tasks were a rarity on statewide tests. Today, they are commonplace, not just in Kentucky and California, but in numerous other states as well. A major reason for the successful incorporation of open-ended writing tasks is the widespread consensus within the education community and the general public that writing skills are crucial and must be assessed directly. Given that clear priority, substantial investments of time and resources were made in the 1980s to support administration of writing examinations and to develop effective scoring paradigms.

In Kentucky, which dropped plans for a mathematics portfolio, the writing portfolio remains an integral part of the statewide high-stakes assessment and accountability system. This is due in part to teachers' dedication to the writing portfolio. It's also due to their preparedness, which results from the state's

strong commitment to providing necessary professional development and logistical support.

In some states, oral presentations may follow the example of writing assessment as a successful assessment innovation. In Oregon, oral communication skills are considered such an important standard that both the state's Certificate of Initial Mastery and Certificate of Advanced Mastery examinations require oral presentations.

Lesson #3: New content areas to be assessed need to fit into the existing assessment frameworks and delivery systems. For C-TAP to finally become operational, it had to be reborn as ACE and blended into the existing statewide student assessment system. This required some significant adjustments. Besides having to conform to multiple-choice and short-answer response formats, ACE assessment items had to meet rules that were originally designed for academic content areas. The previously cited example of ethics items on the ACE Health Careers examination is one illustration of how content in career-technical areas may not easily conform to the rules governing assessments in academic content areas. Delivery system differences must also be considered. For academic examinations, it may be effective to use district-level test administration coordinators to disseminate test information to school sites. But this may not be the best way to reach career-technical teachers and classes because district assessment coordinators do not typically interact with the career-technical departments at their schools. Low initial ACE participation was due, in part, to the fact that in many districts, word of the test did not filter down from district administrators to career-technical education departments and teachers at the school sites.

The ACE example has implications for core academic content areas as well. As more and more states begin to include social studies and science into high-stakes assessment systems, adjustments must be made to align instruction, assessment design, and assessment delivery systems. While many mathematics and English/language arts teachers are accustomed to aligning their instruction to state standards in preparation for assessment and working with district assessment coordinators on the logistics of test administration, most social studies and science teachers have no such experience. Before rolling out high-stakes assessments in new academic content areas, groundwork needs to be laid in both substantive (e.g., alignment of standards, instruction, and assessment) and logistical (e.g., coordination of teachers with district assessment staff) areas.

Lesson #4: There is a need to generate more expertise at the local level for developing and selecting assessments that complement the statewide system. In this era of high-stakes statewide assessment programs, districts' need for staff with solid assessment expertise has never been greater (Rabinowitz & Ananda, 2001). This is true regardless of whether a local district or school plans to implement its own student assessment system to augment the state's assessment system. Simply stated, the higher the stakes in the statewide assessment system, the greater the need for local districts to be informed and critical consumers of tests and test results. There must be sufficient local assessment-related capacity to use assessment data in making decisions.

Furthermore, because the design and implementation of assessment innovations, such as performance-based assessments, are often relegated to the local level, districts must be able to help create innovative assessment systems that meet technical requirements and that complement rather than replicate any existing state testing. Unfortunately, states also need to increase their assessment capacity,

which puts states and locals in the awkward situation of vying for the same limited pool of assessment expertise.

As local education agencies seek to expand their capacity for making the best use of state assessment results and for building their own assessment systems, states must support that effort. The attention commanded by statewide assessment and accountability systems has the potential to overshadow or stifle local assessment initiative. To encourage local initiative, states must find meaningful ways to incorporate local assessments results into their statewide accountability systems.

Lesson #5: Innovations must be fully researched and developed before they are implemented into a statewide assessment and accountability system. We do our schools and students a disservice by incorporating assessment innovations before they are ready to carry the weight of accountability. Premature incorporation of innovation has led several states to scrap the innovative aspects of their systems entirely, delay implementation of innovations, or delay the time when results will count for accountability purposes.

We recommend two avenues for fostering assessment innovation. One approach is for the state to introduce and support innovations at the local level. For example, the Utah State Office of Education promotes and disseminates performance-based assessment models for potential use at the local level. A second approach is to introduce assessment innovations at the state level, but eschew higher stakes. The Vermont writing and math portfolios are examples of statewide assessment innovations in a low-stakes context. Such strategic introductions and phase-ins of assessment innovations are necessary to ensure the integrity and full utility of state assessment and accountability systems.

Conclusion

Despite the difficulties many states are experiencing in their quest for innovations to make assessments more meaningful, fair, and efficient measures of student learning, the future holds significant promise. This study argues that two conditions are essential for effective implementation of assessment innovations in high-stakes systems: (1) a strategic phasing-in of innovations with initial implementation at the local level and (2) establishment of a solid infrastructure (including necessary professional development, teacher comfort level, political will, etc.) throughout the state.

As illustrated in the two cases above, the assessment innovations of the 1990s focused heavily on performance-based assessment. Although, as noted in this paper, the full promise of performance-based assessment has not been realized in statewide assessment systems, performance-based assessment does play a targeted or limited role in many state systems. For example, essays and other constructed-response tasks are now commonplace in many systems. In science, laboratory performance tasks are gaining popularity for statewide end-of-course examinations. Despite the many setbacks to incorporating performance-based assessment in statewide systems over the last decade, it still holds potential as a powerful tool to enhance and assess student learning.

Moreover, the lessons learned from attempts to implement performance-based assessment, as exemplified in the case studies above, generalize to any significant assessment innovation. This includes computer-based and on-line test administration, scoring, and reporting, one of the most visible innovations in the field today (Rabinowitz & Brandt, 2001). Unless these lessons are heeded, we are likely to experience the same types of missed opportunities and

incomplete implementation with computer-based and on-line assessment innovations that we did with performance-based assessment. To survive the scrutiny that attends assessment in a high-stakes environment, innovations must be incorporated in reasonable, incremental, and purposive steps. If this can be accomplished with computer-related assessment innovations, the implications for large-scale assessment would be profound.

References

Evaluation and Training Institute. (1997). *Evaluation of the career-technical assessment program (C-TAP).* Final report submitted to the Sacramento County Office of Education.

Kirst, M. W., & Mazzeo, C. (1996). The rise, fall, and rise of state assessment in California: 1993–96. *Phi Delta Kappa, 78,* 319–323.

Koretz, D. M., & Barron, S. L. (1998). *The validity of gain scores on the Kentucky Instructional Results Information Systems.* Santa Monica, CA: Rand.

Mitchell, R. (1992). *Testing for learning.* New York: The Free Press.

Rabinowitz, S. N., & Ananda, S. M. (2001). *Balancing local assessment with statewide testing: Building a program that meets students' needs.* San Francisco, CA: WestEd.

Rabinowitz, S. N., & Brandt, T. (2001). *Computer-based assessment: Can it deliver on its promise?* San Francisco, CA: WestEd.

Stecher, B. M., Rahn, M. L., Ruby, A., Alt, M. N., & Robyn, A. (1997). *Using alternative assessments in vocational education.* Berkeley, CA: National Center for Research in Vocational Education., University of California.

Wiggins, G. (1993). *Assessing student performance: Exploring the purpose and limits of testing.* San Francisco, CA: Jossey-Bass Publisher.

WestEd, a nonprofit research, development, and service agency, works with education and other communities to promote excellence, achieve equity, and improve learning for children, youth, and adults. While WestEd serves the states of Arizona, California, Nevada, and Utah as one of the nation's Regional Educational Laboratories, our agency's work extends throughout the United States and abroad. It has 16 offices nationwide, from Washington and Boston to Arizona, Southern California, and its headquarters in San Francisco.

For more information about WestEd, visit our Web site: WestEd.org, call 415.565.3000 or, toll-free, (1.877) 4-WestEd, or write:

WestEd
730 Harrison Street
San Francisco, CA 94107-1242.

This report was produced in whole or in part with funds from the Office of Educational Research and Improvement, U.S. Department of Education, under contract #ED-01-CO-0012. Its contents do not necessarily reflect the views or policies of the Department of Education.