Emerging Educational Standards of Performance in the United States

CSE Technical Report 437

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

Robert L. Linn
CRESST/University of Colorado at Boulder

August 1997

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
Los Angeles, CA 90024-6511
(310) 206-1532

The work reported herein was supported 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.
The findings and opinions expressed in this report do not reflect the policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

EMERGING EDUCATIONAL STANDARDS OF PERFORMANCE IN THE UNITED STATES

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

Robert L. Linn
University of Colorado at Boulder
National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Introduction

This paper will describe the role of standards and assessments in the present cycle of educational reform of the United States. Its major sections include a brief description of the American educational system and a reprise of the key elements of the educational reform undertaken in the last decade, including the ascendant role of standards and new forms of assessment, particularly their evolving defining and articulated purposes. We will conclude with the ways in which performance standards, as variously defined, are practically implemented. Illustrating with particular case studies of high-profile performance standards activities, we will project a set of problems that need sustained technical and policy attention for their solution.

The U.S. Educational System

It is incontestable that the U.S. educational system is complex and defies neat summary. It is a mixture of independent and hierarchical policy and operational organizations. Paradoxically, its national character resides in its historical investment in decentralization. The U.S. Constitution assigns educational responsibility principally to the states, and each of the 50 states has developed its own particular flavor. Without exception, each state has a chief education officer, either elected at large or appointed to serve. Although often a nonpartisan office, candidates for the position of chief state school officer may be supported openly or in thinly disguised ways by one or the other of the two major political parties. The portfolio of the chief education officer may focus exclusively on the public schools from kindergarten through

secondary school. In some states, responsibility also may include postsecondary education, libraries, museums, or other less formal educational venues. The policy-making group in each state guiding the commissioner or superintendent (as the chief officer is usually termed) is the state board of education. This group, usually a combination of public representatives with educational, business, academic, or professional credentials, may be elected or appointed. In either case, they may be chosen from the entire state or to represent particular jurisdictions. Clearly the choice of election or appointment influences the chief officers' and the state boards' perception, timing, and locus of accountability. Boards may be augmented with *ex-officio* members.

The chief state officer serves as the head administrator of a department of education, staffed by professionals with responsibilities to implement policies approved by their state board of education and their state legislatures. State boards and departments of education provide for additional public participation by the appointment of standing commissions to deal with issues of curriculum, teacher certification, or higher education policies. These commissions generally include strong representation from educators and academics as well as from the public sector. State legislatures have also created special-purpose commissions and *ad hoc* review groups to deal with particular proposals for reform. There is considerable variation in the degree to which individual state departments of education balance their agenda between initiation of reform and compliance. The policies of each of the state education agencies reflect traditions in the roles of state and local educational authorities and are influenced by factors such as the vast differences in size and composition of the school population, from tens of thousands to many millions of children, or from culturally homogeneous to enormously diverse in language and economic backgrounds.

In every state but Hawaii, the state educational system, its board of education, chief officer, and bureaucracy oversee and operate interactively with local school districts. In Hawaii, the state and local education agencies are coterminous. In the United States, there are more than 15,000 local school districts, a factor of 300 over the number of state authorities. Local school districts have jurisdiction over elementary and secondary schools, although some may also include preschools and two-year postsecondary institutions. As

do the states, local school districts dramatically differ in size, from one or two schools to mega-districts like Los Angeles or New York City, serving 700,000 to one million students, thousands of teachers, and hundreds of schools. Local school districts mirror state organizational structures. Policy is set for them by school boards made up of public representatives. Most of these school boards are elected by the public every two or four years, with staggered terms for continuity. School board elections are often hotly contested; recall elections to depose a board member who has offended a large constituency are not uncommon; and many board members, although receiving nominal compensation, devote considerable energy to their tasks.

These boards have the responsibility for budget, curriculum, operations, and accountability policy. Local school districts must operate within the guidelines and legislatively enacted codes of their states. In the last 20 years, assessing the comparative influence of local versus state school authority, a shift has occurred between the relative influence of state and local school agencies. Although responsibilities for day-to-day educational services remain local, the increasing share of local educational costs borne by states has resulted in centralization of curriculum and assessments. The power first shifted from local school boards to the state departments of education. Yet, as education has become for many states the largest single budget item, educational policy has been seized as a principal issue by the elected state legislatures and the states' overall executives, the governors. Large local school districts represent exceptions to the trend of centralization of power to states. Because they may have strong and direct lines of communication with state legislatures and governors, they may negotiate particular issues or dispensations from certain statewide mandates.

The federal government has limited authority in education matters in the U.S.; nonetheless, it has had significant influence on educational practices in the states, districts, and schools. There is a current debate about whether this influence is positive or intrusive. The federal government implements special programs authorized by Congress to contribute to the educational opportunities of disadvantaged children and special populations, including those with limited English proficiency, those with physical and mental disabilities, and children with other identified needs. These programs contribute significant resources to participating states. In crafting the laws, in

implementing guidelines for local operation of these programs, and in determining the means to evaluate the success of these efforts, the legislative and executive arms of the federal government have a means to set expectations for the groups targeted by the legislation. In fact, their efforts have had impact on all students in the U.S. One particular example was the requirement that a special federal program be evaluated by using standardized tests, administered on a pre and post basis, and reported in terms of national norms. As a result, many school districts adopted a single standardized test for all their students and created public expectations of annual reporting of growth through these measures.

A second, much less expansive way in which the U.S. government influences local public education is in its support of research and dissemination functions. Periodically, agencies of the federal government have supported the design of new teaching materials, teacher training opportunities, and technical assistance. They also support research on fundamental educational processes, such as learning, teaching, assessment, and organization. For example, in the last 30 years or so, the U.S. government has supported a network of educational research and development activities, supplemented by technical assistance service agencies. Organizations participating in research and development do not promote a particular version of the educational reform agenda, but rather provide information and options intended to improve the conceptualization and support of learning, the practice of teaching, and the design of evaluation and governance models. These research organizations function in universities or as independent nonprofit or profit-making entities.

U.S. political realities have made national, coherent programmatic action in support of education very difficult. For the most part, federal financial support has been given to the states with very few guidelines or conditions. Governors, state boards, local educational authorities, and politicians on the right and on the left seek to guard against the danger they perceive in the control of local education agendas from a distal, central source. Through the availability of marginal but discretionary resources, federal programs have reached more than 70% of the schools in the country. But in general, the federal role of education is to identify important priorities and to provide resources to states and localities for implementation. All federal programs

must operate in a tradition that preserves the prerogatives of states and local school districts. Naturally, there are contentions about the location of boundaries among prerogatives.

Although the U.S. government does not control formally the processes of schools in the United States, that is not to say that other national agents have little influence. In fact, because most of the instructional materials, such as textbooks, are commercially developed, there is in fact a *de facto* menu of curriculum embodied in the books and materials that are sold by commercial firms. In general, the process for curriculum marketing involves the presentation of options by commercial textbook publishers to the state board of education or to its delegated body for review at particular grade levels, for example, elementary school, and for particular subject matters, for instance, mathematics or history. Usually states will make primary and alternate choices. Local school districts generally make their selections within these approved choices, for which they typically receive financial incentives for compliance. Because economic considerations preclude creating separate sets of text materials for each state, in practice, the decisions of a very few large states with centralized curriculum selection processes set the boundaries for what is available for other states. A similar situation exists in the area of student achievement testing. A handful of test publishers have a limited number of achievement measures in their inventories. Although these are more frequently modified under contract to meet a particular state need, in fact the majority of achievement tests used in U.S. schools as measures of general system monitoring are commercial products. These commercial tests and textbooks serve as centralizing agents and an important, although indirect, nationalization force in U.S. educational enterprise.

A final background note important to understanding the character of American education is the precept of equal educational opportunity. Although in practice never yet fully realized, at rhetorical, political, and value levels, Americans have long believed their schools should provide educational opportunities for students of all economic, linguistic, and cultural backgrounds. Although approximately 25% of American students advance to a four-year postsecondary educational experience, U.S. policy has been directed to assuring that ethnic and economic backgrounds are reasonably well represented in secondary school graduation and in admission to postsecondary

schools. To that end, both sanctions and incentives have been employed. For example, when pass rates on tests required for secondary school diplomas discriminated against students of African American background, legal challenges were posed to the state. The Office of Civil Rights, for example, has made formal inquiries into the acceptance-denial rates of Asian students at the university level, raising the question of overrepresentation. Obviously, these contradictory legal challenges could not result in simultaneous rational policies; instead, they indicate the tension and concern for fairness and equality in U.S. educational matters.

This commitment to equality, in the light of these conflicts, has in part led the U.S. to avoid *de jure* approaches to early decisions for students about their likelihood to succeed at academic postsecondary institutions. Therefore, most students complete a general academic program in secondary school, one that has been strengthened in the numbers of required courses in mathematics and English language. The equal opportunity commitment is also responsible for policies that encourage children in all educational authorities to complete a full 12- or 13-year educational program, culminating with the high school diploma. This orientation has been informally supported by the lack, until very recently, of secondary education options leading to early careers in the work force rather than college admission.

Yet, the realities of the schools change as conflicts in U.S. society surface. Increasing opportunities for women and the present U.S. economic environment have led to their greater participation in the work force. As a result, schools and other institutions are expected to expand their participation in the overall education of children. Greater career opportunities for women have also reduced the proportion of highly educated women entering the teaching field and have led to a need to improve approaches to teacher recruitment, training, professional development, and sustenance. As a result, program changes for teacher education have been made in higher education institutions and in the credentialing standards of states. Teachers' salaries have been raised, and efforts have been made to improve their work environments. But to date, weak economic growth has left some of these goals unmet, and others likely to be undone. States differ greatly in their level of teacher compensation, their career advancement patterns, and in work environments, such as class size (ranging, for instance, from 22 to 40

students). Teacher unions differ by locality in the extent to which they initiate support or resist various proposed educational changes.

As in past epochs of high immigration and low economic growth, segments of the public have developed some resistance to compensatory or equal opportunity approaches. The public schools, especially in some states with high proportions of minority students, are under increasing scrutiny. Significant political movements have been mounted to provide public support for private schools using voucher or similar programs. At the heart of this challenge are two related concerns: (a) that educational programs are being simplified to meet the needs of students of different cultural, linguistic, or economic background than the majority student; (b) that the schools themselves are organized in such a way to protect and sustain their own bureaucracies as opposed to educating children. These issues have driven the educational establishment to seek substantial reform, undertaken at a time when financial resources are scarce. Furthermore, reviews of existing policies of equal opportunity are underway more broadly in the society. For instance, affirmative action policies are being reconsidered at all levels of government, with concern raised by critics that goals to increase the participation of segments of the population in the work force have created a new kind of unfairness that should be remedied. Similar proposals to revisit the types and kinds of social services available to immigrants attempt to redefine what equal opportunity means.

U.S. Educational Reform and the Development of Content Standards

Almost since its inception, the U.S. public education system has been subjected to scrutiny and repeated calls for improvement and change. Yet, the system has been remarkably adaptable to the changes in geography, demography, and ideology that have characterized the development of its goals. Two distinguishing features of the U.S. setting may provide an important context to understanding its educational reform efforts and prospects. First, the area of education is less likely to be considered the province of expertise, when compared with other public endeavors, for instance, the delivery of health care or the criminal justice system. On an absolute level, respect for U.S. precollegiate educational institutions is not high, and, in fact, colleges and universities are also treated with increasing

skepticism. Teachers' knowledge and competence are often publicly challenged, although there are some visible efforts to identify and acknowledge excellent teaching in both public and private recognition ceremonies. Lay opinion, expressed through local media and governance structures, and correspondingly held public values, has frequently had greater impact in influencing the course of educational reform than has the expert judgment or experience of teachers or scholars.

Second, public interest plays out in the importance of law and litigation as means to direct the education system. Legislation developed from special interests at state levels has resulted in a bewildering number of rules and regulations. These rules cover topics from the number of minutes of instruction in particular subject matters, to insurance provisions, to procedures to deal with armed students. Some regulations address topics to be taught in schools, topics that have been introduced by particular advocacy groups. These requirements generate controversy because they often include areas outside of typical academic pursuits, such as AIDS prevention or personal health. Schools or systems must document compliance with certain regulations, generating microbureaucracies and administrative costs beyond those needed for actually teaching students. Second, litigation has been used as a major form of redress to resolve issues of authority as they intrude on privacy and equal opportunity in the schools. These issues are subject to suit because they are guaranteed by the U.S. Constitution and its amendments. Challenges are raised on behalf of individual parents, students, teachers, and/or other school personnel, or by professional or advocacy organizations, and are initiated by members of every political stripe. Because of the frequency of use and activism of the legal process, many school districts and state education agencies maintain defensive postures. Potential legal recourse is always a consideration before the initiation of new policies and programs. Both legislatively imposed constraints and the threat of legal action serve to limit the flexibility of educational planning and operations.

It is against this backdrop that the current prospects for U.S. educational reform must be analyzed. In the 1980s, following the release of the study *A Nation at Risk* (1983), a series of parallel reforms began to be enacted. One set focused on strengthening the content requirements in schools, resulting in raising the number of courses required for graduation in particular subject

areas. A second wave of reform focused on the teaching profession, particularly on the areas of recruitment, development, and training. A third major change involved the "restructuring" movement, essentially approaches to the devolution of authority and responsibility for school programs and operations to the individual school site level. Devolution worked paradoxically to reduce the impact of local school districts, while simultaneously strengthening the impact of state-level mandates. In the late 1980s, some states embarked on a more systematic course of educational reform. For example, in California and Connecticut, reform focused first on the topic of clarifying content to be learned in schools and resulted in statements called frameworks in particular subjects or common "cores" of learning. These frameworks were to specify broadly the content and knowledge expectations for students and to serve as guidance for related educational functions, including text book and materials selection, teacher professional development, and evaluation of students.

Some of these reforms, in Texas, Arkansas, and Kentucky, were the products of leadership somewhat outside the normal educational channels of superintendents, school boards, and local school districts. The leadership of governors or of the legislatures promoted educational change, articulating a trade-off of greater flexibility, and more resources for higher accountability. The new focus on student competency differed from past reforms that emphasized minimums or the least common denominator of achievement for all students. Instead, the new reforms emphasize high standards of expectation for all students.

In deeper background, private and quasi-governmental reform initiatives were also developing. The National Board for Professional Teaching Standards was created to establish performance standards and assessments for excellent teachers. Researchers were promoting models of teaching, learning, and school change that emphasized a more interactive and constructivist approach to instruction, less homogeneous grouping of students, and more focus on support mechanisms for individual students, including processes for extending the school day, providing homework centers, and other afterschool enrichment options.

Major actors in the reform discussion were members of the business community, in their local, state, and national roles, who promoted reform as the means to restore U.S. economic strength.

In 1989, at an historic meeting on education attended by the governors and the federal administration, a set of National Education Goals was articulated, later endorsed by President Bush in 1990, and proposed as legislation in 1991. One element of this legislation created a council jointly appointed by the governors, the administration, and by Congress, to consider whether national standards and national assessments were desirable and feasible. The deliberations of this council resulted in recommendations supporting the development of state-by-state standards and assessments on a voluntary basis. Of particular concern during this deliberation were two issues: (a) how to assure that high standards and challenging assessments improved the performance of disadvantaged students; and (b) the preservation of state and local authority for educational matters.

In support of goals of the *Raising Standards for American Education* document (National Council on Education Standards and Testing, 1992), the federal government supported the development of content standards (a variation on curriculum frameworks) starting in a number of subject matter areas: history-social studies, science, English language arts, geography, and civics. Following on the widely praised model developed in mathematics by the National Council of Teachers of Mathematics, subject matter specialists and teachers were encouraged to delineate the content standards expected of students. Additional groups lobbied to have national standards developed in their fields, since such standards would be required for inclusion in school programs. This inclusion decision would clearly have impact on undergraduate and graduate programs in the subject matters at the college and university level in terms of areas in which teachers would need strong undergraduate preparation. The first group of subject matter standards are now beginning to be delivered and scrutinized. For example, the standards produced in history were widely criticized for underrepresenting the roots of Western civilization in the American democracy and were repudiated by a vote in the U.S. Senate. Attacks have appeared on the idea of standards themselves with disengagement from the politically bipartisan agreement in Charlottesville, Virginia.

In many states, reform that centers on standards and new kinds of assessments is underway. It is clear that many states are struggling with the problem of making coherent choices among competing standards within and between subject matters. Because of their scope, any one set of subject matter standards, if taken seriously, could easily consume all available student time. Secondly, the subject matter focus of these standards conflicts with another stream of reform: efforts to make school learning more interdisciplinary and more applicable to tasks in real, out-of-school settings, where knowledge from a variety of venues is integrated. Third, in part stimulated by the business community and national panels on problems of the work force, the emphasis on subject matters by standards may not directly serve the needs of the majority of American students not focused on full-time college or university studies. Comparable standards have been set for the development of work force skills, although the procedures for their integration in school curricula have not been thoroughly worked out.

Nonetheless, led by its governor and state education executives, virtually every state is engaged in a systematic rethinking of its educational offerings, of its means of supporting teachers to reach new goals for students, and of its processes for assessing student accomplishments.

Assessments

The U.S. has had a highly developed system for measuring student achievement, and this system has been used extensively in school systems, at state levels and at the national level for a variety of purposes. The purposes include general system monitoring, program or policy evaluation, individual placement and certification, and diagnosis and remediation. For the most part, testing and assessment needs have been met by a set of commercial companies who develop, sometimes administer, and score the tests, and report results to clients. For the most part, many of these measures have been developed to assure that these processes can occur economically. In addition, the history of test development in the United States has created expectations for high technical standards of validity and reliability. These are usually achieved through the application of statistical models dependent upon wide sampling of content and normative comparisons of students and schools. For the most part, these requirements have been met by standardized multiple-choice tests.

The U.S. penchant for educational litigation earlier noted has reinforced the need for tests that are technical defensible, particularly in their fairness to members of disparate populations.

Test Design Options

Technical defensibility is one reason that U.S. educators have used multiple-choice tests as their preferred measures almost without regard to the purpose of the test. Multiple-choice designs also were cost sensitive, for they were cheap to prepare and score. Although examinations requiring students to write essays or prepare projects were sometimes used by teachers to give students marks, open-ended assessments were not often used for publicized accountability purposes. Constructed, open-ended assessments were a part of the early National Assessment of Educational Progress (NAEP) in the 1970s. NAEP was originally conceived both to be a reporting mechanism to monitor national educational attainment and to exemplify and lead the nation's schools in new approaches to assessment. Yet, over the years, needs for efficient information about attainment drove out the assessment leadership role of NAEP. Only recently has it moved from substantially a multiple-choice or constrained test to an open-ended set of measures.

The movement to use more open-ended assessments grew in the 1980s, particularly in the area of student writing. Many states developed or contracted the development of writing assessments, which asked students to prepare essays rather than to edit or evaluate given pieces of work. Although vocally opposed by some members of the testing establishment, the argument for validity (measure writing if that is the goal) overcame objections. Considerable efforts in research and practice have created technical standards for the design of writing tasks and scoring rubrics. The experience in writing set the stage for considering an expansion of open-ended assessments in other subject matters for use in large-scale or accountability situations. Widespread dissemination of the British experiments in the late 1980s and early 1990s also influenced a broadening of testing approach in the U.S. Examples of European school-leaving examination questions were circulated and encouraged U.S. educators to adopt more open-ended testing approaches. The history of performance assessment in the U.S. military, typically used to assess procedural knowledge or strategy in simulated or on-the-job environments,

also contributed to the growing credibility of measures of performance. Furthermore, some educational researchers in the U.S. charged that the low standards of the curriculum were in part caused by the lack of challenge in multiple-choice measures.

In sum, the reforms focusing on higher educational expectations, on new approaches to teaching and learning, and information on assessment from outside sources converged and stimulated an unprecedented degree of agreement among constituencies for educational change: Business leaders, policy makers, teachers and administrators, and researchers have all been vocal in support of assessment reform, even if some may have only vague ideas of what they are actually supporting. Now, at least 16 states have some form of performance-based assessment on a statewide basis. Others have joined in consortia to develop new assessments and 90% of the states have plans to change their assessment systems. There remains some strong suspicion in segments of the public that open-ended assessments are a way to relax rather than raise expectations. Parents worry that these newer (for the U.S.) forms of assessment are subjective and any answer will do—a fear that connects to the skepticism about teacher quality. Establishing credibility for these assessments will minimally depend upon the creation and publication of obviously challenging measures and providing unassailable evidence of their technical quality.

Despite the need for credibility, a number of technical concerns about performance assessment remain, particularly where results of their use will have dramatic consequences for individual students and schools. One troublesome dilemma is the trade-off between adequate sampling of subject matter and the constrained time and resources available for testing. Subject matter sampling decisions interact with the inferences derived from the measures, because different schools may very well teach different aspects of subject matter, and restrictive subject matter representation could result in mismatches between test and instructional content. That these assessments will actually be sensitive to educational interventions is a matter of faith. When assessments have broad boundaries and very general scoring criteria, for example, elaborated versus basic, it is just as possible that individual talent is the major source of variation as the effects of schooling. Because of the striking diversity of student backgrounds in some states and cities, an additional set of

problems exists. For example, is it possible that such assessments can be customized for use with students of non-English-speaking backgrounds and simultaneously maintain the quality of measurement? How will accommodations or adaptations affect the ability to compare the results of different groups?

Assessment, however it occurs, clearly operationalizes the meaning of content standards or curriculum frameworks. It is a key element in understanding the intentions of educational reforms and sends a signal to school practitioners about what is valued. How performance standards fit into the relationship between content standards and assessment is not yet clear.

Performance Standards

We have described the American educational system thus far in order to give some understanding of why current practices in the identification of performance standards are as they are and the options and choices our nation faces in the future.

An essential attribute of any accountability-focused educational system is the way in which one determines success or infers the need to invest differently in educational services. Performance standards serve as a system to divide performers into those who have attained or have yet to attain desired outcomes. These performance standards may also serve as the framework for the public reporting system for educational accomplishment. Reports are promulgated about the average performance of students in a particular subject, or the numbers who have met prespecified levels of accomplishment. If content standards or curriculum goals and targets tell us what is to be learned, and assessments allow us to measure learning, performance standards are to provide guidance about how much or how well students have learned. They are the operational statement of expectations.

Looking at performance standards from a narrow, operational perspective, we see that three major technical approaches have been taken in their formulation. The first two of these take a linear approach starting from the articulation of goals or content standards. Defining performance standards follows the statement of content standards; performance standards may be verbal descriptions of attributes of desired skills at a particular age or proficiency level. For example, the statement that proficient writing is

elaborated with detail and uses a clear line of argument or progression would be an example of descriptive performance standards. The descriptive performance standards are intended to influence the design of the assessment tasks and the approaches used for their scoring. They also may directly imply reporting categories as well.

A second approach to performance standards begins as well with goals or content standards but, forgoing verbal description, operationalizes their meaning in a set of assessment tasks. Variation in the challenge of the tasks themselves serves to illustrate expectations for students, as do model performances that exceed, satisfy, or fail to meet standards. A third form of performance standard involves a preexisting item pool intended to measure the general articulated goals. With this approach, performance standards may take the form of identifying quantitative "cut-scores" or levels of attainment. This approach assumes that the test items measure the same general dimension, so that more accurate, right answers connotes greater competence. In the U.S. considerable research has been undertaken to develop approaches to this sort of performance standard setting. Tests using extensive, open-ended tasks, measuring very different dimensions of content may be ill-served by this cut-score approach to performance standard setting. In these cases, performance may need to be set in terms of qualitative characteristics, categories calling for similar and dissimilar knowledge and skills. Reports of performance in these categories may provide more useful information than the assumption of a linear scale and different quantitative standards of performance.

Some performance levels have focused on different types of quantitative goals, specifying proportions of the group intended to meet particular standards. "Eighty percent of students should master 75% of the items tested" and "failure on no more than one aspect of a multistep task is acceptable for 90% of the students" represent the sort of performance standards often used for the design and evaluation of U.S. training programs in the business and military sectors.

Nonetheless, these technical definitions are far beyond the thoughts of many charged with setting performance standards. In many states, at the present time, special commissions on performance standards have been appointed by legislatures and governors. These consist of broad representation

of the public, parents, policy makers, school leaders, the business community, and teachers. Most of these groups are struggling with the rudiments of selecting general goals or content standards rather than any precise notion of real standards of performance. Most are using descriptive approaches that generally sketch the characteristics or type of content they expect students to have learned by particular points in their schooling, for example, by the end of elementary education. In practice, then, the term performance standard has a loose definition, perhaps assigning different content standards to age ranges, but stopping well before qualitative or quantitative standards of performance have been articulated.

In the next section, we are going to describe experiences with the setting of performance standards drawn from three very different contexts. The first will describe the practices used in setting cut-scores for the Advanced Placement Examinations, voluntary examinations for college-bound students, offered on a national basis annually. These examinations are drawn from a fairly explicit set of curricular expectations, and students take courses to prepare for them specially. A second example comes from the experience of setting performance standards for National Assessment of Educational Progress (NAEP). The third describes the standard setting processes of states committed to using performance-based assessments.

Advanced Placement

The Advanced Placement (AP) Program provides a mechanism for secondary school students to take college-level courses and receive college credit from a large number of colleges and universities in the United States based on their performance on AP Examinations. The program is sponsored by the College Board and is open to any secondary school that chooses to participate. Each course offered by the program includes a course description that provides recommendations regarding the content and skills to be taught and an examination that is tied to the description. AP Examination results are scored under the auspices of the College Board and reported directly to colleges at the student's request.

Most AP Examinations consist of a multiple-choice section and a free-response section. Although the weights differ from one subject matter examination to another, a composite score is obtained as a weighted

combination of the two section scores. The weights for the two sections are reported for the 1986 subject matter examinations in the AP technical manual (College Entrance Examination Board, 1988, Table 4.12). With the exception of Studio Art, which was 100% free response, the relative weights range from 30% multiple choice and 70% free response for History of Art to 75% multiple choice and 25% free response for Music: Listening and Literature. In most cases, the weights are proportional to the time allowed for each section, and the most common weights are 50:50.

Composite scores have ranges that vary from one subject matter examination to another and from one year to another for the same subject. The composite scores are not reported directly, but are used to determine AP grades. The reported AP grades are the scores that are used in college decisions to award college credit and are the scores that are intended to be comparable from year to year.

The AP grades are reported on a 5-point scale. The "grades are: 5, very well qualified; 4, well qualified; 3, qualified; 2, possibly qualified; and 1, no recommendation" (College Entrance Examination Board, 1988, pp. 31-32). Performance standards on an AP examination are established through a judgmental process that determines the four grade boundaries (i.e., the performance standards) on the composite score scale. The AP technical manual provides the following description of this judgmental process:

> Specifying the four grade boundaries is not a simple mechanical process. It can neither be assumed that a given AP Examination is just as difficult as the corresponding examination in a previous year nor that the candidate group is equally strong. Therefore, the Chief Readers select the grade boundaries anew each year for their respective AP Examinations. The choice of each boundary is a judgment based on evidence (College Entrance Examination Board, 1988, pp. 32-33).

Educational Testing Service staff provide statistical evidence, assistance, and advice, but the Chief Reader is responsible for setting the boundaries. Four types of evidence are presented for use by the Chief Readers: (a) distributions of grades from previous years; (b) statistical summaries of candidate performance on multiple-choice items that are repeated from a previous edition of the examination; (c) a "listing or roster of scores on each section, subsection, and individual free-response question for a sample of candidates at

each composite score level" (p. 33); and (d) results of studies of the comparability of AP and college grades. Chief Readers may also rely on their own experience from review of candidates' free-response answers.

Personal experience in grading and profiles of performance on parts of an examination for various composite scores (the third type of evidence) could be used for most any effort to establish performance standards for a given assessment. The first two types of evidence (previous grade distributions and statistics based on repeated multiple-choice items) are potentially relevant to continuing assessment programs but do not help in the establishment of initial performance standards. Even for a continuing assessment program, the amount of weight given to these first two types of evidence in determining the standards depends upon judgment about the likelihood that the current assessment is more or less demanding than previous assessments and the likelihood that the achievement of the examinees taking the current assessment is better or worse than that of examinees from previous years.

The fourth type of evidence (results of comparability studies of AP and college grades) is perhaps the most unusual and is the only source of evidence that attempts to use an external criterion to inform the standard setting judgment. Studies of the comparability of AP and college grades are done for all new AP Examinations and are repeated periodically for established AP Examinations (CEEB, 1988). In a typical study of the comparability of AP and college grades, a portion of an AP Examination is administered to college students who are enrolled in courses for which a college would normally award credit based on AP Examination results. Statistical analyses of the results provide a comparison between existing AP grading boundaries and the grades assigned independently by college professors.

The National Assessment Governing Board's Achievement Levels

The National Assessment of Educational Progress (NAEP) was initially designed in the mid 1960s to provide dependable measures of the progress in student achievement in the United States on a periodic basis. Although the fundamental objective of measuring educational progress for the nation has not changed, NAEP has undergone a variety of modifications during the quarter century since the first National Assessment was administered in 1969.

Of particular relevance to the focus of this paper are changes in the ways in which NAEP results are reported.

The early assessments reported results at the individual item level. Released items (referred to as exercises) were presented along with the percentage of students who answered the item correctly or performed the task successfully. Average percent-correct statistics were used to summarize the results, but those statistics have a number of limitations, including the requirement that identical sets of items be used when comparisons are made from one assessment to the next or among results for different grades (see Phillips et al., 1993, for a more complete discussion).

Starting in 1984 the NAEP results were reported in terms of scale scores for each content area. With the exception of the Writing Assessment, the scales were developed using item response theory. Results were summarized using means and selected percentile points. "Anchor Points," which correspond to selected scores on the scale corresponding to the combined age group mean and one or two standard deviations from the mean, were also used. The anchor points were accompanied by descriptions of what students scoring at or near those points were able to do on the assessment (see Phillips et al., 1993).

The first use of performance standards-based reporting of NAEP results was with the 1990 mathematics assessment when the National Assessment Governing Board (NAGB) developed its first set of "achievement levels." Among other responsibilities, the 1988 legislation that created NAGB charged the board with the task of "identifying appropriate achievement goals for each age and grade in each subject area to be tested under the National Assessment" (Public Law 100-297). NAGB interpreted this part of the legislation as a mandate to set performance standards for NAEP, which were labeled achievement levels. Although other approaches might have been used to fulfill this responsibility (e.g., establishing targets in terms of the percentage of students scoring above existing anchor points), the approach of setting standards that establish what students should be able to do was consistent with several other national efforts that encouraged the establishment of standards of student achievement (e.g., the National Education Goals Panel, 1992; the National Council on Education Standards and Testing, 1992).

NAGB provided policy definitions for three achievement levels labeled Advanced, Proficient, and Basic. Starting with those policy definitions, panels of judges developed operational definitions of the levels and identified cut-scores for the levels based on ratings of the NAEP item pools for each grade. The work of the panels of judges was used by NAGB to establish three cut-scores that divided the NAEP proficiency scale into four regions: Below Basic, Basic, Proficient, and Advanced.

Three separate evaluations of NAGB's initial effort to set achievement levels for the 1990 mathematics assessment all reached negative conclusions regarding both the process and the outcome of the undertaking (Linn, Koretz, Baker, & Burstein, 1991; Stufflebeam, Jaeger, & Scriven, 1991; U.S. General Accounting Office, 1993). Among the major criticisms were the following: (a) the achievement level descriptions and associated exemplar items did not adequately coincide with actual performance of students scoring at a given achievement level; (b) there was a lack of evidence to support the validity of interpretations invited by the achievement level descriptions; (c) the NAEP item pool was not adequate to measure the advanced levels; (d) the judgment process was too demanding for raters; and (e) the standards were overly dependent on the particular sample of judges. It should be noted that NAGB made adjustments in response to some of the early criticisms (e.g., the lack of coherence of levels from grade to grade noted by Linn et al., 1991), which included the assembly of additional panels of judges. Because of technical difficulties, NAGB also re-established achievement levels for mathematics in 1992. In addition, achievement levels for the reading assessment were established for the first time in 1992.

The 1992 achievement levels also proved to be controversial despite efforts by NAGB and its contractor, the American College Testing Program (1993), to improve the standard setting process. The first evaluation of the 1992 effort (Burstein et al., 1993), which focused exclusively on the adequacy of the descriptions of the 1992 levels in mathematics, concluded that the descriptions and associated exemplar items did not adequately characterize what students scoring at a given level can do. The lack of correspondence between achievement level descriptions and exemplar items in mathematics, on the one hand, and actual student performance on the assessment, on the other, led to changes in the selection of exemplar items and the descriptions used for

the 1992 reading achievement levels. Nonetheless, both the 1992 mathematics and reading achievement levels were judged to be unacceptable in two other evaluations (National Academy of Education, 1993; U.S. General Accounting Office, 1993).

The National Academy of Education (NAE), while strongly affirming "the potential value of voluntary national standards that exemplify challenging curricular and performance expectations" (1993, p. xxiv), was critical of many aspects of process used to set the achievement levels and of the resulting levels and descriptions. The NAE evaluation concluded that the method used as well as "other item judgment methods are fundamentally flawed" (p. 132) and therefore recommended that NAGB and the National Center for Education Statistics "discontinue reporting by achievement levels as used in 1992" (p. 132). As is obvious from the title of its report, "Educational Achievement Standards: NAGB's Approach Yields Misleading Interpretations," the GAO evaluation was also highly critical (U.S. General Accounting Office, 1993).

The NAE recommended that achievement level results be published separately from the main NAEP results in research and development reports while a longer-range effort is undertaken to establish performance standards on NAEP. NAGB and its consultants (Cizek, 1993; Kane, 1993) have attempted to rebut many of the criticisms of the achievement levels, and the board remains firmly committed to reporting achievement level results.

In an effort to resolve conflicts regarding achievement levels, NAGB and the National Center for Education Statistics sponsored a conference on standard setting in October 1994. A variety of perspectives were represented at the conference. There was considerable agreement on the desirability of establishing standards, the need to make them understandable to a wide variety of audiences, and the importance of broad representation in the formation of panels of judges. There was also strong support for the creation of descriptions that are valid reflections of what students who meet a given standard of performance know and can do. There was less agreement, however, about the best way to accomplish that end and about the criteria that should be used in evaluating the degree to which the goal was achieved. For example, should students who meet a given standard be able to do all, two-thirds, or a majority of the tasks that are implied by the description?

The other large issue that remained unresolved concerns the judgments that panels of judges are asked to make. The traditional approach, which involves judgments regarding expected performance of students who meet a given standard separately for each individual item or task on an assessment, remains the preferred approach of one group of experts, but continues to be rejected by another group that argues that judgments need to be based on a consideration of a complete record of performance (e.g., the pattern of responses to all assessment items and tasks) for actual students. The latter position is reflected in the report of the National Academy of Education (1993), while the former position continues to be adhered to by NAGB and its consultants.

State Assessments

A number of states have adopted or are in the process of developing standards-based reporting procedures for their statewide assessments. This introduction of standards-based reporting coincides with the movement toward a greater reliance on performance assessments and is consistent with the national press for the development of student performance standards (see, for example, National Education Goals Panel, Goals 3 and 4 Technical Planning Group on the Review of Education Standards, 1993).

Performance standards divide the continuum of student attainment on an assessment into two or more levels of achievement. The number of levels varies from state to state. In the 1970s and 1980s many states introduced minimum-competency testing programs that required the establishment of a minimum standard of performance for the award of a high school diploma. Some states still have minimum competency or certification testing programs where the emphasis is on a single, pass-fail standard. The recent emphasis on standard-based reporting, however, usually has more than two levels.

Kentucky, for example, has set three performance standards in each content area resulting in four levels of achievement that are labeled Distinguished, Proficient, Apprentice, and Novice (Trimble, 1994). Each labeled performance level on the Kentucky Information and Reporting System (KIRS) is accompanied by a definition that is intended to communicate what students achieving at that level are able to do. The levels were defined in response to the Kentucky Education Reform Act (KERA) of 1990 requirement

that schools be held accountable for increasing the proportion of "successful" students. Hence, the system began by defining the Proficient standard to correspond to a level of performance where a student would be considered successful, which, among other things, is intended to indicate that a student has the "skills necessary to function in a complex and changing civilization" (Kentucky Education Reform Act, 1990). Students who perform at the Novice level demonstrate few, if any, of the skills and understandings defined by the Proficient standard. The Apprentice level is intermediate between Novice and Proficient and students at that level need to provide "tangible evidence of 'making progress' toward the Proficient standard" (Trimble, 1994, p. 47). Distinguished, the highest level, "was established to recognize the accomplishments of a small percentage of students who exceed even the Proficient standard" (Trimble, 1994, p. 47).

Maryland has set four standards that yield five levels of achievement for the Maryland School Performance Assessment Program (MSPAP). The numbered levels, where 1 is the highest and 5 is the lowest, are each described by a list of activities that students scoring at that level on the assessment are able to do (e.g., at level 1 in mathematics students are, among other things, able to "make predictions using basic concepts of probability in abstract settings" while students scoring at level 4 are able to "describe relationships among data in a chart/table," Westat, 1993, pp. 16-17).

Results on the California Learning Assessment System (CLAS) are reported in terms of six performance levels, with 6 denoting the highest level of performance and 1 the lowest (CTB/McGraw-Hill, 1994). As is true for the MSPAP, the CLAS levels are not labeled, but each is defined by a description of what students who score at that level can do.

The procedures used by states to establish performance standards vary in their details but share a number of common features. Usually, the process has two key components: (a) the definition of performance levels, and (b) the mapping of scores on the assessment into the performance levels. Both components rely on judgments of one or more panels that are assembled specifically for that purpose. Although described separately here, the tasks of developing definitions and mapping assessment results into performance levels are sometimes done iteratively through three or more steps (e.g., define levels, map assessment scores into levels, revise level definitions).

Definitions. Definitions of performance levels typically start with a predetermined number of levels. Panelists may be asked to review a state curriculum framework or content standards as well as actual assessment tasks. They may begin with policy statements regarding the standards that are provided by the state legislature or the state school board. As was implied above, the legislative requirement of defining "successful" students led to an initial focus on the Proficient standard in Kentucky, for example. A higher standard, Distinguished, was added to provide recognition for exceptional achievement, and a lower standard, Apprentice, was added to acknowledge students who were making progress but had not yet achieved the Proficient level, and with three standards, the fourth performance level, Novice, was defined by default.

Starting with a broad framework and possibly a general policy-level description of standards, panels of judges typically are asked to develop and refine definitions of performance levels. The emphasis is usually on defining performance that students "should" achieve and hence is forward-looking rather than purely normative in nature. States vary in how closely tied the definitions of levels are to the actual assessments. In Maryland, for example, panelists were instructed to include in the description of a given proficiency level only student outcomes or aspects of performance that were actually assessed (Westat, 1993, p. 14). This procedure is intended to avoid problems that have been encountered in other assessments where the descriptions included aspects of performance that are not assessed and therefore may not validly describe what students who achieve a given level actually know or are able to do.

Mapping. In some assessments, the levels correspond naturally to the score categories defined by scoring rubrics used for open-ended student responses. This direct correspondence between a scoring rubric and the performance levels used for reporting is, perhaps, most common for writing exercises. In some state writing assessments, a single prompt may be used at each grade where students are assessed. Student essays are scored according to state scoring rubrics, which typically have four, five, or six score levels. With results from a single performance task scored on 4- to 6-point scale it is natural to simply use those scores to correspond to performance levels.

Scoring rubrics are also sometimes translated directly into performance levels where multiple tasks are included in the state assessment. For example, although each student responded to only one prompt, several different prompts were administered to different samples of students in the 1993 CLAS reading assessment. Student responses were all scored on a 6-point scale and those six score points were judged to correspond directly to the six CLAS performance levels by the committees assembled to develop descriptions of performance levels and map student performance on the assessment into those levels (CTB/McGraw-Hill, 1994).

Where assessment scores have more possible values than performance levels, some process is needed to map scores into levels for purposes of reporting. Most often the mapping consists of setting a series of cut-scores that divide the scores into as many ranges as there are performance levels. Scales based on item response theory models such as the ones used in each grade and content area for the MSPAP provide a direct way of arranging student performance and assessment tasks on the same scale. As was done for MSPAP, judges can be asked to identify regions of the scale where the associated items correspond most closely to the definitions of student performance for a given level. By a process of averaging, those judgments are readily converted to cut points on the scale which then map the scale scores for students into the proficiency levels.

With a small number of open-ended assessment tasks the judgment process may involve a review of exemplar student performances corresponding to each possible score. Such a process was followed in Kentucky where content committees reviewed sets of student responses to three tasks per student. The sets of student responses corresponded to each of 10 possible score points (3 to 12 for the sum of scores of 1–4 on three tasks per student), and judges were asked to convert each of the scores on the 10-point scale into one of the four Kentucky proficiency levels (National Academy of Education, 1993).

A slightly more complicated mapping procedure is illustrated by the conversion of student assessment results in mathematics to performance levels on the 1993 CLAS. As part of the matrix sampling design, each student responded to one open-ended mathematics problem and seven multiple-choice items on the 1993 CLAS. Rather than trying to place both open-ended tasks and multiple-choice items on a single scale, judges were asked to review the 32

cells of a 4 by 8 matrix of possible student outcomes (4 possible score levels on the open-ended task by the 8 possible scores of 0 to 7 for the number correct score on the multiple-choice items). For each of 16 possible combinations of open-ended tasks with sets of multiple-choice items used in the 1993 CLAS matrix sampling design, judges were asked to map each of the 32 cells into the six performance levels used by California.

Standard-based reporting of results by states is a relatively recent phenomenon. Thus, it is too early for any comprehensive evaluation of the approach. Certainly, none of the state standard setting efforts has undergone the kind of stringent evaluation to which the NAGB standard setting effort for NAEP was subjected. Two things are clear at this stage, however. First, the standards-based reporting of results is popular among state policy makers and directors of assessment programs. Second, there is strong encouragement for increased use of performance standards in both federal and state legislation. States are encouraged to develop performance standards by the Goals 2000: Educate America Act that was signed into law by President Clinton in the spring of 1994. That encouragement is reinforced by requirements in the subsequently enacted the Improving America's Schools Act that authorizes $7.4 billion for the Title I compensatory-education program targeted at schools serving children from low-income families. Schools receiving Title I funds will be required to report progress in terms of the percentage of students who are meeting advanced, proficient, or partially proficient levels of performance. Hence, an increased emphasis on the use of performance standards by states can be anticipated.

Performance Standards and Inferences From Assessment

The key issue in the design of performance standards is the degree to which the standards are set with an eye to the purpose for which the assessment is designed. If the purpose of the assessment is to operationalize a quota in which limited spaces are allocated to the best students, then the details of performance assessments may not be of much interest, save the estimation of the likelihood of misclassifying individuals. If the performance standards are set to provide targets for educational systems to reach, then it is important that the attributes used to define the highest levels of performance are those that the system can address. Very little research has been conducted

to validate performance standards, particularly those that include specification of student response attributes. The validation of standards raises deeper questions of assessment purposes. Are the purposes of assessments fundamentally to make predictions about students' success in the workplace, or in higher or further education? When educational systems serve increasingly diverse students, questions of the postmodern type abound. Should all students reach the same standards? Are there comparable but different standards students of different backgrounds or different aspirations might meet? How is comparability established to assure that systems of multi-tiered quality do not develop? How does one push for national educational improvement in a federal system, with strong local and site control of goals, curriculum, and testing programs? How do we manage the tension between homogeneity and adaptation when we have different approaches, strategies, and belief systems in the public?

Finally, we know little about how to set up procedures to yield quality and valid performance standards. Who should participate? What content, school or technical background should participants have, if any? At the present, approaches to setting performance standards vary considerably, from those that simply attempt to arrive at consensus to those that systematically provide various kinds of information about students, normative performance, and model answers to the standard setting group. As noted above, the relationship is yet unclear between the design of scoring rubrics, created in part to achieve high technical standards, and the development of reporting categories, which need to have public credibility. Should rubrics and standards be simultaneously developed? Does it matter which comes first?

Last, there is a set of issues around the use of performance standards to make judgments about institutions as well as individuals. Although different viewpoints clearly exist, there is some sentiment to provide contextual information to aid in the understanding of levels of attained performance. Context can be introduced by statistical adjustments of scores, to show how students would have performed had they had similar backgrounds, school or classroom sizes, etc. The statistical adjustment approach is criticized because it masks, and usually overestimates the measured performance of the least well performing students in the system. Other strategies for contextualizing performance include the concurrent reporting of collateral data regarding

school resources, student backgrounds, and information about students' opportunity to learn assessed material. This set of information may be useful not only to help the target audiences to understand why some institutions outperform others, but to guide improvement of performance.

Without doubt, the U.S., with its decentralized approach to education, has a complex path to negotiate if it is to emerge with high-quality performance standards that meet educational and public criteria for effectiveness, utility, and credibility.

References

Advanced Systems in Measurement and Evaluation, Inc. (1992). *Kentucky Instructional Results Information System 1991-1992 technical report*. Dover, NH: Author.

American College Testing Program. (1993). *Setting achievement levels on the 1992 National Assessment of Educational Progress in mathematics, reading and writing: A technical report on reliability and validity*. Iowa City, IA: Author.

Burstein, L., Koretz, D. M., Linn, R. L., Sugrue, B., Novak, J., Lewis, E., & Baker, E. L. (1993). *The validity of interpretations of the 1992 NAEP achievement levels in mathematics* (Technical Report). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Cizek, G. (1993). *Reactions to National Academy of Education report, "Setting performance standards for student achievement."* Unpublished manuscript, University of Toledo, Toledo, OH.

College Entrance Examination Board. (1988). *The College Board technical manual for the Advanced Placement Program*. New York: Author.

CTB/McGraw-Hill. (1994). *California Learning Assessment System, 1993: Preliminary technical report*. Monterey, CA: Author (Draft, February, 25).

Kane, M. (1993). *Comments on the NAE evaluation of the NAGB achievement levels*. Unpublished manuscript, University of Wisconsin, Madison, WI.

Kentucky Education Reform Act of 1990. KRS 158.645.

Linn, R. L., Koretz, D. M., Baker, E. L., & Burstein, L. (1991). *The validity and credibility of the achievement levels for the 1990 National Assessment of Educational Progress in mathematics*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

National Academy of Education. (1993). *Setting performance standards for student achievement. A report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: An evaluation of the 1992 achievement levels*. Stanford, CA: Author.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform. A report to the nation and the Secretary of Education*. Washington, DC: U.S. Government Printing Office.

National Council on Education Standards and Testing. (1992). *Raising standards for American education: A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American people*. Washington, DC: U.S. Government Printing Office.

National Education Goals Panel. (1992). *The National Education Goals report, 1992: Building a nation of leaders*. Washington, DC: U.S. Government Printing Office.

National Education Goals Panel, Goals 3 and 4 Technical Planning Group on the Review of Education Standards. (1993). *Promises to keep: Creating high standards for American students*. Washington, DC: Author.

Phillips, G. W., Mullis, I. V. S., Bourque, M. L., Williams, P. L., Hambleton, R. K., Owen, E. H., & Barton, P. E. (1993). *Interpreting NAEP scales*. Washington, DC: National Center for Education Statistics.

Public Law 100-297. (1988). *National Assessment of Educational Progress Improvement Act* [20 U.S. Code Sec. 1221e-1].

Stufflebeam, D. L., Jaeger, R. M., & Scriven, M. (1991). *Summative evaluation of the National Assessment Governing Board's inaugural effort to set achievement levels on the National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board.

Trimble, C. S. (1994). Ensuring educational accountability. In T. Guskey (Ed.), *High stakes performance assessment: Perspectives on Kentucky's education reform* (pp. 37-54). Thousand Oaks, CA: Corwin Press, Inc.

U.S. General Accounting Office. (1993). *Educational achievement standards: NAGB's approach yields misleading interpretations* (Report No. GAO/PEMD-93-12). Washington, DC: U.S. Government Printing Office.

Westat, Inc. (1993). *Establishing proficiency levels and descriptions for the 1992 Maryland School Performance Assessment Program (MSPAP)*. Rockville, MD: Author.