Recent National Trends in Assessment

Over the past two years there have been a few significant activities and projects at the national level that seem to be redefining the nature of both student and teacher assessment. As one might expect, the majority of activities have focused on student assessment and methods to raise the level of student performance. Indirectly, these activities have implications for teacher training and, consequently, the nature of teacher licensure and certification testing. The following review of the activities in student assessment and description of a national project on teacher certification will reveal three consistent trends in testing: (1) the increased use of tests to drive educational reform, (2) an emphasis on measuring higher-order cognitive skills and complex competencies, and (3) a shift toward performance-based assessment methods to measure those skills and competencies.

A chronology of the activities in student assessment is outlined in Table 1.

New Standards Project

The New Standards Project was developed early in 1991 at the Learning Research and Development Center (LRDC) at the University of Pittsburgh and the National Center on Education and the Economy (NCEE) by Drs. Lauren Resnick and Mark Tucker (LRDC & NCEE, 1991; Resnick, 1991; Tucker, 1991). This project was supported by grants from the MacArthur Foundation and Pew Charitable Trusts. The major goal of the project was to create a national examination system that “will be part of an educational system structured to reward
students who work hard to meet a clear mastery standard and provide school staff with the resources and incentives they need to help all students meet that standard” (LRDC & NCEE, 1991, p. 1).

**Table 1**

**Chronology of National Activities in Student Assessment**

<table>
<thead>
<tr>
<th>Date</th>
<th>Activity/Project</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>d. Compromise bill?</td>
<td></td>
</tr>
</tbody>
</table>

This examination system has three characteristics that distinguish it from current testing programs: (1) it is syllabus driven, such that teachers would be expected to help students prepare for the exams based on criteria known to both students and teachers; (2) it uses the “three Ps”—projects, portfolios, and performance tests—as the examination formats; and (3) it relies on tasks determined locally or by
clusters of states and districts, although a national standards board would oversee and contribute prototypes of tasks and technical assistance. National standards would be derived from products of states (or clusters of states and districts) or from some form of national anchor examination.

In the summer of 1991, ten states participated in a cross-state scoring workshop on writing samples. The results of a study comparing judgments of student writing across states were recently published (Linn, Kiplinger, Chapman, & LeMahieu, 1992).

**National Council on Education Standards and Testing (NCEST)**

In June 1991, the National Council on Education Standards and Testing was created by Public Law 102-62, with Governors Roy Romer (CO) and Carroll Campbell (SC) appointed as co-chairs. The purpose of the Council was to provide advice on the desirability and feasibility of national education standards and a system of voluntary national tests. The goal of any such system would be to foster good teaching and learning, as well as to monitor performance.

This charge to the Council touched off a wave of political controversy on the merits of a national test. Gov. Romer was one of the most outspoken advocates of such a test. As the “national test movement” gained momentum, the current standardized testing program came under attack. “Standardized test-bashing” was commonplace, with politicians and members of the public pushing for a new form of assessment that would break the mold of traditional types of testing. Throughout the remaining months of 1991, a shift toward “authentic” or “performance-based” assessment was evident. This “break-the-mold” assessment would replace all existing forms of assessment, which were judged to have failed.

Coupled with this national test and performance assessment rhetoric was the demand for “world-class standards” that would require high levels of performance on challenging subject matter in English, mathematics, science, history, and geography in grades 4, 8, and 12. Since traditional testing formats were considered inadequate for measuring higher levels of cognition (e.g., higher-order thinking skills), it seemed most appropriate that a performance-based testing
system would inspire students to learn at, and teachers to aim for, these higher levels—rather than the “minimum competency” level which dominated the 1970s and 80s.

**America 2000**

In September 1991, the National Education Goals Panel, chaired first by Gov. Romer and then by Gov. Campbell, issued its first report, the *National Education Goals Report: Building a Nation of Learners* (a.k.a. *America 2000*). Six national goals were specified along with defining objectives. The report set the year 2000 as the deadline. The six areas addressed were (1) readiness for school, (2) high school completion, (3) student achievement and citizenship, (4) science and mathematics, (5) adult literacy and lifelong learning, and (6) safe, disciplined, drug-free schools. These topics provided the springboard for the development of national content standards in the workings of NCEST.

**Raising Standards for American Education**

In January 1992, NCEST submitted its report, *Raising Standards for American Education*, to Congress and the National Education Goals Panel. Of the six goals presented by the Panel, NCEST focused on how to measure progress on goals 3 and 4 only. The NCEST Report recommended that the National Education Goals Panel and a new National Education and Assessments Council:

- provide for and coordinate the development of voluntary national standards for content, student performance, and system performance;
- provide for the development, by states, of school delivery standards;
- certify content and student performance standards as world class;
- provide for the development of a system of assessments for individual students which is aligned with national standards;
- develop assessments to monitor the performance of programs and systems consistent with national standards;
- provide research and development for “break-the-mold” assessments;
- issue quality guidelines for the development of assessments;
• ensure the technical merit (i.e., validity, reliability, and fairness) of assessments;
• certify that assessments are aligned with national standards; and
• establish procedures and criteria for achieving comparability among assessments.

The assessment component emphasized the use of the National Assessment of Educational Progress (NAEP) to monitor overall progress in meeting national goals and the voluntary participation of state and local educational agencies in developing, acquiring, and administering new “break-the-mold” performance assessments which measure individual student achievement against national content and student performance standards. There was no mention of a national test anywhere in the Report. That idea was replaced with the notion of a multilevel national examination system.

Federal Legislation

Soon after the NCEST Report was issued, the U.S. Senate passed S.2, which contained the entire report.

In response to the NCEST Report’s recommendations, the U.S. House Subcommittee on Elementary, Secondary, and Vocational Education conducted a series of hearings over the next two months. The cautions expressed by witnesses in these hearings, along with the February 1992 Office of Technology Assessment report entitled Testing in American Schools: Asking the Right Questions and a counter-proposal by the National Council on Measurement in Education (NCME) for a nonpartisan coordinating entity and list of technical assessment standards, led to a more conservative posture on the issues of national standards and an assessment system.

The result of this cautionary input and several caucuses in April and May 1992 was H.R. 4323, the Neighborhood Schools Improvement Act, which represented the Subcommittee’s response to America 2000 and the NCEST Report. This Act was more comprehensive than the President’s proposal and it encouraged states and local school districts to improve their schools through broad-based, coordinated “systemic” reform. It established the National Education Goals Panel and a Technical Review Committee to oversee the development and certification of voluntary national content and school delivery standards. This bill provided for expanded research on performance
assessment and for models for assessing national standards for mathematics, the only national content standards currently in existence. Additionally, the bill authorized the National Academy of Sciences (1) to evaluate the national delivery standards, the model mathematics assessments, and assessment research conducted through the Office of Educational Research and Improvement; and (2) to identify criteria for evaluating the technical adequacy of the assessments and whether they are aligned with national content standards. H.R. 4323 passed the House in August 1992.

Given the major differences between the House and Senate bills and the change in administration, it is impossible to predict whether a compromise version will emerge from this process.

These important activities at the national level have implications for teacher assessment, particularly with regard to professional training for instructional improvement and using the new types of assessment. However, while these student assessment projects and bills in Congress were being developed, an independent teacher certification project was undertaken by the National Board for Professional Teaching Standards. The characteristics of this project and the role of performance assessment methods in teacher certification are examined next.

**National Board for Professional Teaching Standards**

The National Board for Professional Teaching Standards, a non-governmental, private, non-profit organization, was established in 1987. The Board has 63 members, consisting of two-thirds classroom teachers and one-third business leaders, local and state officials, and others. Its goal is to create a system of voluntary, advanced certification for teachers.

The mission of the Board is (1) to establish high and rigorous standards for what accomplished teachers should know and be able to do, (2) to develop and operate a national voluntary system to assess and certify teachers who meet these standards, and (3) to advance related educational reforms for the purpose of improving student learning in American schools.
This mission is to be accomplished in three phases: (1) establishing the policies that will guide its work, (2) translating the characteristics of advanced teaching into specific standards that will apply to each certificate, and (3) field testing the system through Assessment Development Labs and a national Field Test Network.

Let’s briefly examine what is planned in each of these phases.

**Phase One**

Phase one involves setting policies to execute the project. The first policy specifies five propositions that define accomplished teaching: (1) dedication to students and their learning, (2) mastery of the subjects they teach and how to teach those subjects to students, (3) effective management and monitoring of student learning, (4) thinking systematically about their practice and learning from experience, and (5) working together with other professionals and parents in the interest of student learning. These characteristics combine critical skills in scholarship with an understanding of child development and effective instruction.

**Phase Two**

Phase two consists of translating these characteristics into specific standards for each certificate and then developing an assessment process to measure them. It is expected that certificates will be offered in 30 fields, each of which designates a specific subject at a particular student level, such as Early Adolescence/Science. Standards committees will be formed to set standards that describe the specific knowledge and performance that best reflect accomplished practice in each field. Each committee will be chaired by a teacher and composed of a majority of skilled teachers. To date, 12 committees have been organized for 12 certificates.

Once the standards are set, the assessment procedures can be developed. This step will be accomplished by Assessment Development Laboratories, which are responsible for designing the methods and materials that will measure a candidate’s advanced skills and knowledge. Through competitive bidding, the Board has awarded contracts for two Laboratories. In January 1992, a Technical Analysis Group (TAG) was contracted to conduct necessary research on technical issues related to the assessments and to assist the
Laboratories in development and field testing. This TAG is based at the University of North Carolina at Greensboro, under the leadership of Prof. Richard Jaeger.

What are the criteria for building this assessment system? It should be administratively feasible, professionally credible, publicly acceptable, legally defensible, and economically affordable. The actual assessment will primarily utilize performance assessment methods. One component will take place at the candidate’s school, where portfolios and classroom observation are used; a second component will occur at an assessment center at a different school site.

**Phase Three**

The third phase of this certification system is field testing, which will be conducted through a Field Test Network of school districts nationwide. Teachers in these sites will test the assessment process for the initial certificates. Field testing will also occur in the Assessment Development Labs. This phase will begin in 1993 and continue until 1998.

**National Board Certification vs. State Licensure**

By way of summary, how do the characteristics of this National Board Teacher Certification compare with those of the current State Teacher Licensure assessments? Five characteristics are examined in Table 2.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>National Board Certification</th>
<th>State Licensure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Target Population</td>
<td>Experienced Teachers</td>
<td>Beginning Teachers</td>
</tr>
<tr>
<td>2. Assessment Scope</td>
<td>Voluntary Nationwide</td>
<td>Mandatory by State</td>
</tr>
<tr>
<td>3. Assessment Domain</td>
<td>Advanced Skills and Knowledge (Scholarship, Child Development and Effective Teaching)</td>
<td>Subject Matter Content</td>
</tr>
<tr>
<td>4. Primary Assessment Method</td>
<td>Performance Assessment</td>
<td>Written Examination</td>
</tr>
<tr>
<td>5. Performance Standards</td>
<td>High Levels of Competence</td>
<td>Minimum Competence</td>
</tr>
</tbody>
</table>
Performance Assessment

Definitions

Before diving into the specific performance assessment methods, I would like to suggest a definition or two of performance assessment. If you review the burgeoning literature on this topic, you will rarely find an explicit definition anywhere. There is a sense that everyone already knows what “performance” or “authentic” assessment means, and that a definition is unnecessary. However, given the rather narrow focus of many of the writings, it might be instructive to consider the scope of the methodology via two definitions. The first definition was proffered by the Office of Technology Assessment in their February 1992 report to Congress:

Performance assessment refers to testing methods that require students to create an answer or product that demonstrates their knowledge or skills. Performance assessment can take many different forms including writing short answers, doing mathematical computations, writing an extended essay, conducting an experiment, presenting an oral argument, or assembling a portfolio of representative work (Office of Technology Assessment, p. 3).

This definition encompasses the numerous methods typically reported in research and projects dealing with student assessment. This seems appropriate, since I would guess about 80% of attention paid to the topic has focused on students, with the remaining 20% on teachers and administrators. For our purposes, the OTA definition is clearly unacceptable. I have broadened the definition to read:

Performance assessment refers to testing formats that require individuals to create an answer or product or to demonstrate physically the completion of tasks that measure their achievement. Such formats include work samples or hands-on-tasks, written and computerized simulations, behavior-based rating scales, oral presentations, and portfolios (Moi, p. 10).

Now we can proceed to review comfortably the specific performance assessment methods used for National Board Certification.

National Board Certification Methods

Rather than study the National Board Certification methods in isolation, let’s view them in the context of the vast array of performance assessment methods developed in industrial/organizational and personnel psychology over the past 30 years. The first column of
Table 3 provides an inventory of those methods used in the performance appraisal of supervisors and employees in business, industry, and the military. To the right are the specific methods identified by the National Board to be used for teacher certification. Many of these same methods have already been incorporated into many licensure tests.

**Match between teacher certification methods and those of personnel psychology.**

What can we conclude about the match between these two columns? The performance assessment methods planned for teacher certification are not new. Conceptually they are virtually identical to those used in personnel psychology. What is different in their application to teacher assessment is the form, based on the different content domains or job analyses of teachers to which the methods are applied. In fact, the assessment center components that are to be developed for teachers have been used with school administrators for at least ten years. Given the similarity between these two columns, it seems reasonable to conclude that we can draw on the corpus of research that has already accumulated on these methods to evaluate the technical issues and practical concerns.

**Mismatch between teacher certification methods and those of personnel psychology.**

In comparing the two columns, it seems obvious that there are several methods used in personnel psychology that have not yet been considered in teacher assessment: (1) numerical, graphic, and behavior-based rating scales to measure behaviors by direct and indirect observation; and (2) verbal, written, and computerized simulations. These methods are certainly worth exploring in view of the heavy emphasis on direct observation of teacher performance.

**Teacher certification methods that are unique to education.**

Constructed-response items, essay questions, and writing samples which can be used in teacher licensure and certification tests have been used in educational assessment since “prehistoric” times (Baby Sinclair, 1992). The written examinations, as part of the assessment center, may employ some of these item formats.
Table 3  
Comparison of Performance Assessment Methods  
Used in Personnel Psychology vs. Those Being  
Considered in Teacher Certification

<table>
<thead>
<tr>
<th>I/O-Personnel Psychology Methods</th>
<th>Teacher Certification Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Numerical and Graphic Rating</td>
<td>Observation Checklist</td>
</tr>
<tr>
<td>Scales</td>
<td>(Classroom Observation)</td>
</tr>
<tr>
<td>2. Behavior-Based Rating Scales</td>
<td></td>
</tr>
<tr>
<td>a. Behaviorally Anchored</td>
<td></td>
</tr>
<tr>
<td>b. Behavior Summary</td>
<td></td>
</tr>
<tr>
<td>c. Behavior Observations</td>
<td></td>
</tr>
<tr>
<td>3. Work Samples</td>
<td></td>
</tr>
<tr>
<td>a. Verbal (or written)</td>
<td>Class Portfolios (Videos, Student</td>
</tr>
<tr>
<td>b. Motor (Manipulative)</td>
<td>Work Samples, Teacher Commentary</td>
</tr>
<tr>
<td>c. Situational</td>
<td>and Artifacts)</td>
</tr>
<tr>
<td>d. In-basket</td>
<td></td>
</tr>
<tr>
<td>4. Simulations</td>
<td>Teaching Situations</td>
</tr>
<tr>
<td>a. Verbal (or written)</td>
<td>(Videotaped) Role Play</td>
</tr>
<tr>
<td>b. Live Performance</td>
<td></td>
</tr>
<tr>
<td>c. Computerized</td>
<td></td>
</tr>
<tr>
<td>5. Assessment Center</td>
<td>Assessment Center</td>
</tr>
<tr>
<td>a. In-basket</td>
<td>In-basket</td>
</tr>
<tr>
<td>b. Leaderless Group Discussion</td>
<td>Group Exercises</td>
</tr>
<tr>
<td>c. Interview Simulation</td>
<td>Structured Interview</td>
</tr>
<tr>
<td>d. Scheduling</td>
<td></td>
</tr>
<tr>
<td>e. Analysis</td>
<td></td>
</tr>
<tr>
<td>f. Management Games</td>
<td></td>
</tr>
<tr>
<td>g. Background Interview</td>
<td></td>
</tr>
<tr>
<td>h. Paper-and-Pencil Tests</td>
<td>Written Examinations</td>
</tr>
<tr>
<td>i. Projective Tests</td>
<td></td>
</tr>
<tr>
<td>6. Appraisal Interview</td>
<td></td>
</tr>
</tbody>
</table>
Technical Issues

Some preliminary conclusions can be made regarding the technical adequacy of these different performance assessment methods (see also Dunbar, Koretz, & Hoover, 1991). These are based on the experiences and research involving these methods in business and industry, the military, and the medical and health professions. The conclusions are presented according to Messick’s (1989) evidential basis of test score interpretation and use and the consequential basis of test score use (social consequences).

Evidential basis

These issues relate generally to construct validity, including content- and criterion-related evidence, but also pertain to evidence for the relevance of the assessment tasks to the specific applied purpose and for its utility in applied settings.

1. Construct Representation: The content representation of performance tasks in relation to the behavioral/job domain may indicate that the performance assessment is too narrow and fails to include important dimensions of the construct. For example, in work samples and other hands-on performance tests, efforts to select the most critical and frequently performed tasks do not preclude the limited sampling of tasks from the domain. The final small number of tasks, by itself, typically underrepresents the construct.

2. Convergent and Discriminant Evidence: Correlations of performance tests with other measures of the same construct, as well as with measures of other constructs, are inconsistent. For work samples, assessment centers, and simulations, convergent and discriminant evidence are needed to support construct interpretations.

3. Task Generalizability: The tasks sampled for a particular performance test often exhibit “content specificity”; that is, examinees’ scores vary across different situational tasks or problems. Generalizability coefficients for tests in a variety of disciplines tend to be moderate and examinee by-task interaction effects have been reported.
4. *Validity Generalization:* Performance tasks and test-criterion relationships may not generalize across settings and times. The tasks used in work samples and assessment centers to evaluate teachers may be situation specific. Situational factors such as teaching style, type of classroom or school setting, and level of motivation can affect performance on the same tasks or job in different settings or at different times.

**Consequential basis**

This type of evidence is based on the appraisal of both potential and actual social consequences of applied performance assessment:

5. *Adverse Impact and Bias:* Systematic differences among genders and ethnic subpopulations have been found on different types of performance tests, for example, hands-on tasks and essay tests. Investigations of construct bias, selection bias, and item bias (differential item functioning) are needed for performance tests just as for traditional objective item formats. Judgmental and empirical evidence must be collected to assess adverse impact.

6. *Scoring Method:* A variety of scoring systems need to be tested with different assessment methods, for example, holistic/global versus analytical/point, weighted item or option, expert or empirically derived aggregate scoring. The type of scoring protocol and the rigor with which it is executed can markedly affect the generalizability of the scores.

7. *Interscorer/Rater Generalizability:* The magnitude of generalizability coefficients varies with the performance assessment method used, scoring protocol, and the extent of scorer/rater training.

**Practical Issues**

Despite all that has been written and attempted with performance assessment methods to date, there remain two prominent concerns: (1) the *administrative time* to execute the methods, whether one-on-one, in small groups, or station-by-station, is very lengthy compared to alternative methods, and (2) the *costs* associated with development, administration, and scoring are extremely high compared to alternative methods.
Sources and Standards

Since much of what is known about performance assessment has been reported in books and journals outside of the field of education, a listing of the most relevant sources has been compiled along with a few educational sources that are reporting the results of performance assessment research and projects (see Supplement).

Considering the number of technical issues that require attention, a set of technical and legal standards is needed to evaluate the quality of the products that are being generated nationwide. A set for performance assessment only does not exist. Standards do exist, but are embedded in the standards documents already well-known to educators. The EEOC Uniform Guidelines are also germane to judging validity issues and, especially, adverse impact. These sets of standards and guidelines are identified in the Supplement.

Conclusions

The preceding review of national activities and projects in student and teacher assessment suggests several common threads or trends in testing: (1) tests are being used increasingly to drive broad-based systemic reform of curricula, instructional methods, instructional materials, and professional development, (2) tests are being designed to measure higher-order skills, advanced levels of competence, and rigorous standards of performance, and (3) a variety of performance assessment methods, such as portfolios and assessment centers, are being administered in lieu of or in conjunction with traditional item formats to measure the higher-level skills. The notion of a single national test was eventually abandoned in the major student assessment projects in favor of a multilevel (e.g., LEA, SEA) assessment system, yet national teacher certification hinges on the single-test approach.

Despite the apparent newness of the performance assessment methods planned for teacher certification, almost identical methods have been employed in business and industry for more than 30 years. While the forms of some of the methods need to be adapted for teachers, research in personnel psychology indicates significant technical and practical problems that have not been overcome. These problems have not restricted their use in the private sector, and it seems unlikely that they will deter educators from applying these break-the-mold methods to teacher assessment.
Notes

1 The author gratefully acknowledges the assistance of Rorena Washington in the preparation of this chapter and the materials accompanying the presentation upon which the chapter was based.

2 The descriptive information on the National Board for Professional Teaching Standards presented in this section is adapted from a presentation on that topic by Jaeger (1992).

References


Berk


**Supplement: Suggested Sources and Standards on Performance Assessment**

**Books**


Journals

Psychology

Journal of Applied Psychology

Military Psychology

Personnel Psychology

Education

Applied Measurement in Education (1991, 4 [4])

Educational Measurement: Issues and Practice

Journal of Educational Measurement (1993, 30 [2])

Standards


