NBETPP Boston College

The Building Blocks of State Testing Programs

National Board on Educational Testing and Public Policy

By Arnold Shore, Joseph Pedulla, and Marguerite Clarke

Lynch School of Education

Boston College

August 2001

Statements Series

Volume 2
Number 4

NBETPP

The Building Blocks of State Testing Programs

Arnold Shore, Joseph Pedulla, and Marguerite Clarke
National Board on Educational Testing and Public Policy

INTRODUCTION

Across the country, states have put into place educational testing programs to assess student learning and to hold educators or students accountable for learning outcomes. At present, most states have mandated programs that test students in several grades; the future will likely see more students tested in more grades. As the growth in testing continues, it becomes imperative that these tests do the least harm and bring the greatest good to the education of elementary and secondary school students. To ensure that the tests meet this criterion, parents, teachers, educational administrators, and policy makers must be actively involved in their construction and use. The purpose of this Statement is to provide a checklist of some of the components of these testing programs and the choices they represent.

In the first part of the Statement, we describe six basic building blocks of state testing programs. In the second, we demonstrate with two simple examples how these building blocks can be put together to form two very different state testing programs. While other building blocks could be added to the list, these six are highlighted because they reflect the main dimensions along which current state testing programs differ, and around which most debate and controversy occur.

Statements Series Editor:
Marguerite Clarke

"As the growth in testing continues, it becomes imperative that these tests do the least harm and bring the greatest good to the education of elementary and secondary school students."

SIX BASIC BUILDING BLOCKS

Not in any particular order, the six building blocks of state testing programs are:

1. **Stakes levels –** the consequences for students, teachers, administrators, schools, districts, or school systems that flow from the results of the test

2. **Performance standards –** the criteria against which academic performance is measured

3. **Information dissemination –** what is made public about the test results, to whom, and when

4. **Involvement of teachers –** the level and degree of teacher input into test questions, the frameworks that organize those questions, the scoring system (or "rubrics") for the test, and the scoring of the test

5. **Technical design and data –** degree of compliance with professional standards for development, review, and use of the test

6. **Range of measures used to assess educational performance –** from the use of a single test (given once or several times) to the use of multiple measures (e.g., teacher grades, portfolios of students' work, the state test) given over a period of time

Let us discuss each building block in turn.

1. Stakes levels

The stakes levels of state testing programs can vary from low (e.g., no observable consequences attached to the test results) through moderate (e.g., public dissemination of test results) to high (e.g., consequences ranging from diplomas withheld to loss of accreditation for schools). As the examples imply, the stakes levels can vary in another way: they can apply to students on the one hand and to "schools" (read: teachers, principals, other school administrators, schools, districts, systems) on the other. Putting the two types of variation together – level of consequences and bearer of consequences – a state testing program could have high stakes for pupils (e.g., a state test determining graduation) coupled with high stakes for schools (e.g., testing outcomes determining salaries, resource allocation, or even operating independence), or low stakes for students and high stakes for schools, and so on.

Figure 1 presents the stakes-level choices that states have made for both students and schools in their current testing programs.

Consequences for Students

Consequences for Teachers, Schools, and Districts

	High	Moderate	Low
High	Alabama Alaska* California* Delaware * Florida Georgia* Indiana* Louisiana Maryland* Massachusetts* Mississippi* Nevada New Jersey New Mexico New York North Carolina Ohio South Carolina Tennessee Texas Virginia* Wisconsin*	Arkansas Colorado* Connecticut Illinois Michigan Pennsylvania West Virginia	Kansas Kentucky Missouri Nebraska Oklahoma* Rhode Island Vermont*
Moderate	Arizona* Minnesota Utah* Washington*	Oregon	Hawaii Maine Montana New Hampshire North Dakota South Dakota Wyoming
Low	Idaho*		Iowa

**Indicates that not all aspects of the program are in place.*

Which stakes level should be chosen – low, medium, or high – for which group – students or schools – depends on the goals involved. Here, as for all the building blocks that follow, the refrain is this: only by being clear about the educational goals of the testing program can one make these choices in an informed way. Also important is the need to consider district, school, and teacher resources as well as the extent to which students have had an opportunity to learn the material tested when creating a timetable for attaching stakes to test results.

Which stakes level should be chosen– low, medium, or high– for which group– students or schools– depends on the goals involved.

2. Performance standards

Most state testing programs measure progress toward the content standards (what you know or are able to do) and performance standards (how well you know it) that are laid out in the state's curriculum frameworks (what is taught, and when; what is learned, and when). When it comes to an actual test, setting the performance standards for the content tested in various grade levels (say, 4th, 8th, and 10th) requires the setting of scores that separate categories of students, for example those who "fail," "meet basic expectations," "exceed basic expectations," and are "proficient." These points of separation are called "cut scores" and setting them requires technical expertise as well as the exercise of judgment (see the National Board monograph "Cut Scores: Results May Vary").

One must recognize that the cut scores chosen for a given test could differ dramatically depending on the goals involved and the cut-score setting method employed. Three possibilities will help illustrate the range of options.

1. The cut scores can be set with an eye toward "real-world" requirements for verbal, mathematical, scientific, or other skills

2. The cut scores can be set low and raised over time with an eye toward improvement that is deemed possible or probable given the educational starting points of the students, the educational resources of the school district, and the educational leadership of the school system

3. The cut scores can be set high to send an unmistakable signal to teachers, administrators, students, and parents that world-class standards of achievement are the goal of the state's educational system

The first option relies mainly on the judgment of subject-matter experts. The latter two rely more on the judgments of policy personnel or politicians. Clarity of educational purpose is necessary for making these choices since the cut scores chosen will have very different repercussions for students, teachers, and schools.

3. Information dissemination

Tests provide feedback information. That said, it is a matter of choice when the information is provided, who gets it, the level of its detail, and the ways in which it is presented, or aggregated, for students, classrooms, schools, districts, or states.

When test scores are released can affect how they are used. For example, state tests are often given in the spring and results announced in the fall. Then, for students who took the tests as, say, 4th graders, teachers may receive the results when the students are 5th graders, a time when they have different teachers and when effective action to help individual students is no longer possible. For schools and districts, too, this timing is unfortunate. Results are reported too late to change curriculum or instructional emphasis until the following fall (when the student who took the test is in the 6th grade). A simple rearrangement to test in fall and report in spring (or earlier) would enable both students and teachers to make timely use of the feedback information the testing program provides.

Tests do not supplant teachers. At their best, they support teachers' work by assessing students' progress.

This choice of timing, and other choices as well, comes down to the intended purposes or goals of the testing program. Who gets what results, the level of detail of those results, and the level of aggregation at which the results are reported all relate to the intended purposes and uses of test results. If the main purpose is to help teachers identify strengths and weaknesses of individual students with respect to attaining the state standards, results must get back to teachers in a timely manner. This would argue for fall testing, rapid scoring and reporting to teachers, detailed diagnostic analysis for each student, and little, if any, aggregation of results at the school-building or school-district level.

If, on the other hand, the main purpose of the testing program is school-level accountability, results for individual students become less important. Although timely reporting is still desirable, spring testing with reporting the following fall may meet the needs of this program. Results would be reported at the school-building level and given to building administrators, district administrators, and State Department of Education personnel. The results of this type of program would have less direct impact on any individual student.

4. Involvement of teachers

Tests do not supplant teachers. At their best, they support teachers' work by assessing students' progress. If state tests are to do minimum harm and maximum good, it follows that teachers must be involved from the start in their development, scoring, and use.

Teacher involvement can vary greatly. At the high end, teachers would help guide test development, scoring, and the timing and use of test information so that it would be most useful for instructional practice and classroom assessment. Indeed, teachers brought into the process may find creative ways to link state-mandated testing across classrooms with teacher-developed assessments within classrooms, to the benefit of teachers and students. For example, teachers' classroom assessments might focus on subject matter covered or specialized skills (e.g., debating skills) emphasized in classroom work, while state-mandated tests might, in complementary fashion, test more general reasoning, communication, and problem-solving skills.

Validity is at the heart of the technical information necessary to support the use of any test.

At the low end of involvement in state testing programs, teachers might act merely as test administrators without a further role in test development, scoring, or use. To elaborate possibilities for ratcheting up their involvement, it would be useful to understand how existing state testing programs affect teachers. In a current study, the National Board is asking teachers how these programs affect teaching and learning in their classrooms, especially what they teach and how they teach it. We are also asking about the effect of the tests on students and the ways schools and school districts spend their time and money. Finally, we are asking how the tests affect the profession of teaching.

5. Technical design and data

Validity is at the heart of the technical information necessary to support the use of any test. By validity we mean the extent to which the implications – or inferences – drawn from the test score are accurate. Information independent of the test is required to validate whether a student who fails the test is also likely to fail the next grade's academic work, for example. Categorizations like "failing" that attach test results to teachers or schools also require evidence of their validity. Thus, if a school is categorized as "excellent" on the basis of test scores, independent data on teacher preparation, teacher instruction, and school leadership needs to be gathered and reported. The reason is simple: individual test scores aggregated to produce high scores for a school could result from what the students bring to the classroom, not what they get there. Family involvement in homework and study, rather than what the school is providing, may account for the results.

While validity is the main technical concern of testing, reliability – the consistency of results over time – is also important. Since all test scores embody some degree of error, categorizations such as "failing" and "needs improvement" could change over time for students or schools, independently of whether they actually improve. To buttress claims of a highly reliable test, data need to show that there are few such changes in categorization. Since results are categorized on the basis of cut scores (see above), program administrators also need to provide data on how the cut scores were set, whether multiple procedures (and perhaps multiple panels of judges) were used, and – since judgment always enters into setting cut scores – who was involved in the process.

The test items need to be screened to ensure that they do not disadvantage any group of students. If multiple forms of the tests are used, technical data should be provided to show how the tests were "equated." The treatment of scores obtained when accommodations are made for special populations (e.g., the visually impaired, the learning disabled) also needs to be described: that is, will these scores be treated as equivalent to or different from scores obtained under standard conditions, and if different, how will they be treated?

Finally, the test data need to be analyzed to help make sense of the results. For example, are poorer school districts performing better or worse than expected, given their resources? Are students of a particular racial or ethnic group performing better or worse than other groups? It is only by identifying differences such as these that we can begin to understand where problems and successes lie and how we can address the problems and spread the successes.

All of this information is needed to interpret results accurately. The guiding principle in reporting technical data is to give sufficient information so that others can make informed decisions about the adequacy of the tests and the inferences made from them. This means that test developers should offer guidance on the ways test results are intended to be used, and – equally important - on the potential limitations of the test, including examples of uses that would be inappropriate.

The guiding principle in reporting technical data is to give sufficient information so that others can make informed decisions about the adequacy of the tests and the inferences made from them.

6. Range of measures used to assess educational performance

On the face of things, the range of measures used to assess educational performance can go from a "low" of a single test to a "high" of multiple measures, including classroom work, the state test (or series of tests), school completion rates, and so on. In general, the assessment of educational performance should not be confined to the low end of the range, especially if the stakes attached to a student's or school's performance are high. That said, choices to be made include the number of measures to be used, their format or characteristics (e.g., portfolios of students' work as compared to standardized multiple-choice tests), and their relationship to different outcomes (e.g., relationship to future academic success versus future work success). There is room for much creativity in bringing together a range of measures to assess educational performance, including the use of a computer-based information system to allow teachers to input multiple measures into an accountability system where they and others can track student and school progress.

In the next section we put together the six building blocks to develop two very different testing programs. Note that these examples are not offered as ideal programs. They serve merely to illustrate how the building blocks we discussed play out under two different policy goals.

BUILDING THE STATE TESTING PROGRAM

Several organizing principles help guide the construction of a state testing program. Since that program represents a testing policy, the following can be helpful in making choices:

- **Goals –** what we wish to accomplish with the testing program
- **Coverage –** whom we will include in the program, and when
- **Cost –** how much we will spend to carry out the program
- **Administration –** how we will organize and implement the program
- **Equity –** what safeguards we will put in place to ensure reasonable and fair treatment of those tested
- **Evaluation –** how we will know what has transpired so that we can improve the program

…testing policy, and the testing programs guided by it, should complement a more comprehensive approach to assessing the performance of a student, a teacher, a school, or a district.

For a testing program to be implemented thoughtfully, fairly, within budget constraints, and with an eye toward future improvement, one must be clear about its goals. Immediately below we will take the first steps of focusing on program goals and their relationship to the six building blocks. In the final analysis – a level of analysis not attempted here – we would run the goals and building-blocks choices against coverage, cost, administration, equity, and evaluation to arrive at a polished statement of the testing policy we wish to implement as a testing program.

As we turn to stating goals and outlining our testing programs, we need to assert the fundamental guideline that testing policy, and the testing programs guided by it, should complement a more comprehensive approach to assessing the performance of a student, a teacher, a school, or a district. We underscore that from a technical as well as an educational standpoint, a single test cannot assess performance adequately, especially when an important decision is to be made. Thus, neither of our examples uses just one measure to assess educational outcomes.

Testing Program I

In the first scenario, the policy goal for the testing program is the following:

> *To determine the level at which students are performing in fundamental subject areas (e.g., mathematics, science, English) and the rate of progress we might reasonably expect of them over a certain period of time.*

Our program goals would be to establish the level at which the current state educational system is operating and to involve teachers, curriculum specialists, professional development experts, and the like in assessing where additional or different efforts (always constrained by money, time, and resources) should be spent. In addition, we would involve educational-measurement specialists not only in establishing baseline information, but in reviewing research and conducting further research to determine how much progress can reasonably be made over what period of time.

In constructing a testing program to implement these operational goals, the building blocks might line up this way:

- **Stakes levels:** set low or moderate
- **Performance standards:** set in two ways – in the early grades, ability to do academic work in the next grade, and in the later grades, ability to succeed in the world of work
- **Information dissemination:** test results made widely available in a timely fashion
- **Involvement of teachers:** close involvement in specifying standards and test items, scoring these items if they are open-ended, uses of data, and timing of information dissemination
- **Technical design and data:** meeting relevant professional standards and guidelines, including timely and open dissemination of information on test design and quality
- **Range of measures used to assess educational performance:** reliance on classroom measures, state testing measures, and amount of resources given to schools

Under this scenario, parents, teachers, educational administrators, measurement experts, and policymakers would deliberate jointly on:

- When to use state tests and when other measures
- How often to assess performance using state tests and other measures
- How to keep teaching and curriculum informed by the best possible information

Note that here state tests play a supporting role, not a dominant role; high stakes would not be part of the testing policy; and widely disseminated test information would be part of an effort to bring parents, teachers, school administrators, and decisionmakers into close conversation about student learning.

Testing Program II

In a very different scenario, let us state the testing policy goal as follows:

> *To measure attainment of world-class standards through a test that is independent of the classroom but connected by content and skills to the work of students in specified grades*

The first operational goal would be to establish the level at which students need to perform in order to meet world-class standards by involving state educational administrators and policymakers, in setting cut scores for acceptable and unacceptable levels of performance. A few administrators and political decisionmakers would then make key testing decisions, especially on when students must reach the designated performance goals.

In this scenario, the building blocks would line up this way:

- **Stakes levels:** set high to send an unmistakable signal that the state will take action that has serious consequences for students, teachers, and schools not meeting performance standards
- **Performance standards:** set high to signal that world-class standards will be required of all students

- **Information dissemination:** school and district test results widely publicized, with districts not meeting the standards placed on notice of possible consequences to operations or resources
- **Involvement of teachers:** at most, teachers involved as test administrators to keep costs down
- **Technical design and data:** difficult to assess compliance with relevant professional standards and guidelines due to the limited dissemination of information
- **Range of measures used to assess educational performance:** reliance mainly on the test, but with classroom assessments and district resources being taken into account when determining consequences

In this scenario, administrators and policymakers would deliberate together on:

- Where to set cut scores
- When to administer tests
- How to relate the political environment in which education operates to efforts to reform education according to a standards-based model

In this scenario, the performance standards are set without regard to baseline information; teachers are not involved in discussions about test development or use; and classroom assessment – teacher-developed tests and teacher-based observations – is very much secondary in important educational decisions such as promotion and high school graduation.

To contrast Testing Program I and Testing Program II overall, the first scenario is essentially a developmental approach to learning and assessment. It involves teachers, parents, and others in determining what is appropriate for the educational development of students and how tests can provide information to teachers and parents for furthering this end. The second scenario is essentially a joint political and educational approach to setting educational standards and assessing educational outcomes. It embodies the notion that political leadership is necessary to get educational systems to respond to the need for improved student performance, and that tests play the key role in reorganizing or reforming education.

CONCLUSION

A state testing program is based on a state testing policy. That policy is made up of building blocks that encompass everything from the technical design of the test to performance standards, and to the consequences, if any, that flow from the results. And although testing policies and programs are multifaceted and complex, their potential implications for students demand that the interested public take an active role in their construction.

This paper has provided a necessary but not a sufficient basis for the involvement of parents, other community members, teachers, administrators, and policymakers in deliberations on tests and testing policy. Communities must come together to review the design of their testing program, to affirm goals for testing programs and the underlying testing policies, and to engage in making educational decisions that use testing in support of student learning.

About the National Board on Educational Testing and Public Policy

Created as an independent monitoring system for assessment in America, the National Board on Educational Testing and Public Policy is located in the Carolyn A. and Peter S. Lynch School of Education at Boston College. The National Board provides research-based test information for policy decision making, with special attention to groups historically underserved by the educational systems of our country. Specifically, the National Board

- Monitors testing programs, policies, and products
- Evaluates the benefits and costs of testing programs in operation
- Assesses the extent to which professional standards for test development and use are met in practice

This National Board publication series is supported by a grant from the Ford Foundation.

The National Board on Educational Testing and Public Policy
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Telephone: (617)552-4521 • Fax: (617)552-8419
Email: nbetpp@bc.edu

Visit our website at **nbetpp.bc.edu** for more articles, the latest educational news, and information on NBETPP.

NBETPP

The Board of Overseers

BOSTON COLLEGE