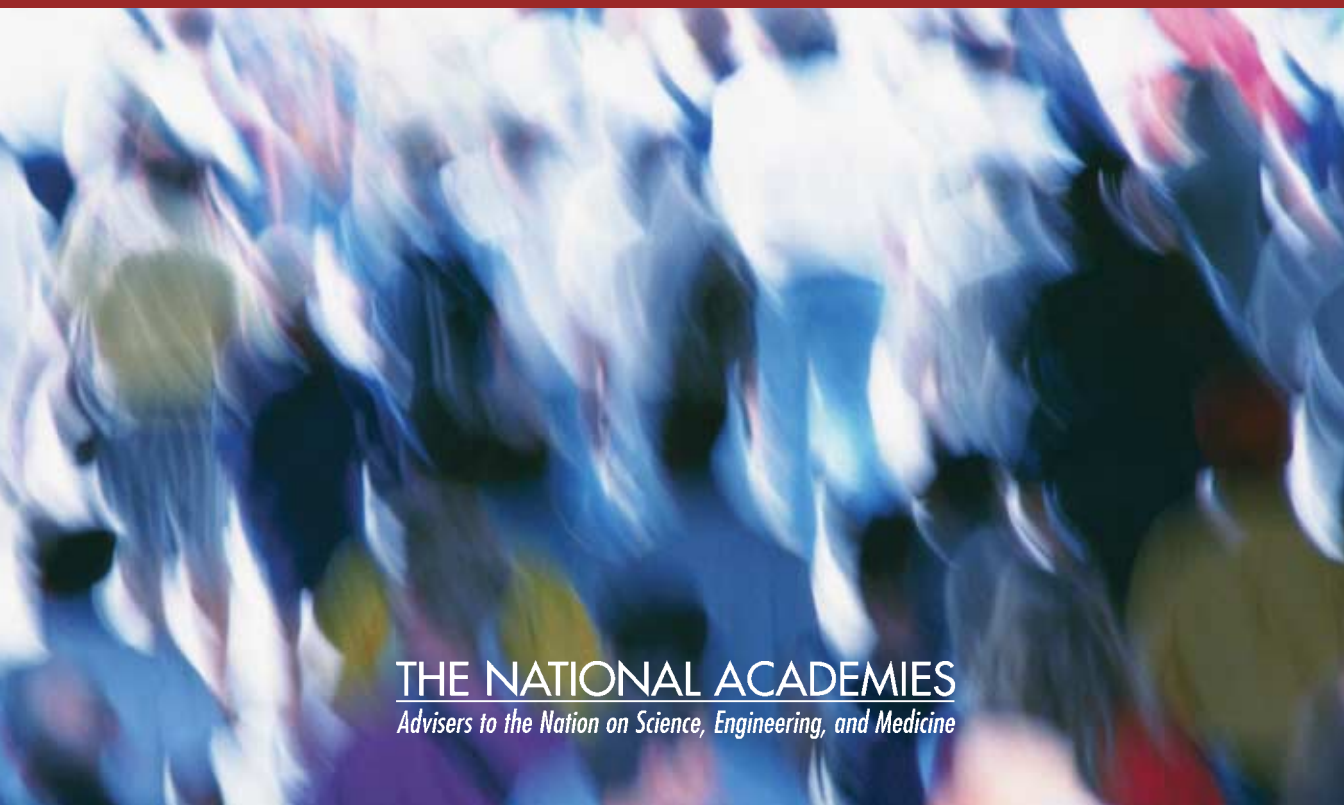
Lessons Learned ABOUT TESTING

TEN YEARS OF WORK AT THE NATIONAL RESEARCH COUNCIL



THE NATIONAL ACADEMIES
Advisers to the Nation on Science, Engineering, and Medicine

The National Research Council (NRC) was formed in 1916 to further knowledge and advise the U.S. government on scientific and technical matters; its "clients" include the U.S. Congress, government departments and agencies, and private foundations. In 1993 the NRC created the Board on Testing and Assessment (BOTA) to apply scientific expertise to critical issues of testing in education, the workplace, and the military. Through this board, the NRC advises policy makers and practitioners about the strengths and limitations of tests, as well as their appropriate design, use, and interpretation.

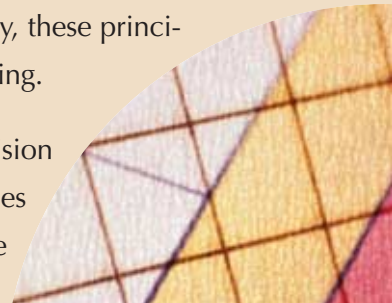
Over the past ten years, this NRC board has explored some of the most pressing issues in assessment, especially in education. The standards-based reform movement of the 1990s, culminating in the federal No Child Left Behind Act of 2001, has greatly increased reliance on testing in education. Through BOTA, the NRC has studied the effects and uses of high stakes tests for students, how to test students with disabilities, and the civil rights implications of tests. It has addressed current trends and controversies about testing for other purposes, such as college admissions, licensing teachers, and adult literacy.

The NRC has also explored innovations in testing that hold promise for the future, such as how advances in the cognitive sciences, measurement, and technology could be applied to develop assessments that provide more useful information about student achievement and support learning. The NRC has advised policy makers on testing programs outside of education as well, such as the redesign of the U.S. naturalization tests.

The NRC's Board on Testing and Assessment consists of experts from a range of disciplines relevant to testing and assessment, including psychometrics, psychology, applied linguistics, statistics, education, economics, law, business, anthropology, sociology, and politics. The board holds several meetings every year, during which members, sponsoring agency staff, and guests discuss current issues in testing and hear presentations by invited researchers and policy makers. Board membership changes on a rotating basis to ensure new perspectives as well as continuity.

The Board on Testing and Assessment convenes committees of experts to work on specific issues and projects. Committee members represent diverse viewpoints, and the committee reports represent the consensus views of leaders in the nation's scientific community. This booklet is based on the reports authored by these committees over the past ten years. These reports, cited throughout, are listed and described at the end of the booklet.

From ten years of work on testing, the NRC's Board on Testing and Assessment has found that several themes arise again and again across a variety of studies. These recurring themes, or lessons, are basic principles in testing that all stakeholders in testing programs should understand. Yet, though widely accepted in the professional testing community, these principles are often neglected in practice and therefore bear repeating.



The lessons presented in this booklet are geared toward decision makers in education who use large-scale tests, particularly ones that carry high stakes for individuals. In this document, the terms "testing" and "assessment" are used interchangeably, to refer to a variety of means for gathering information about student or examinee performance. "High-stakes tests" are those used to make decisions with important consequences for individuals or institutions, such as whether a student will receive a high school diploma or whether a school should be restructured. Many of the principles also apply to the design and use of tests in other education contexts, such as the kinds of tests teachers use in their classrooms. Although most of the lessons stem from the NRC's work on testing in education, many of the same principles also apply to testing in other contexts, such as the workplace.

This booklet is by no means a comprehensive guide to the proper use of tests in all situations. A number of existing documents, most notably the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council of Measurement in Education, 1999), fulfill that purpose. Rather, the goal in this booklet is to highlight some of the major messages from the NRC's work and to direct the interested reader to the relevant published reports for more in-depth coverage of these issues. The lessons are grouped under four broad topics: uses, design, consequences, and public understanding.

USES

Tests are used in education to measure what students know and can do. When used correctly, they can be a relatively objective and efficient way to gauge student achievement. *High Stakes* (1999) recounts that beginning with the introduction in the mid-19th century of written examinations given to large numbers of students, standardized tests have served as an instrument for accomplishing a variety of policy purposes, including determining the types of instruction individual students receive, shaping the content and format of that instruction, and holding schools and students accountable for their performance.

Standardized tests are now believed to be one of the most powerful levers that elected officials and other policy makers have for influencing what happens in local schools and classrooms. This trend has culminated in the federal No Child Left Behind Act, which holds schools and states accountable for increasing test results, and applies sanctions to schools that do not show adequate results. At the same time, those concerned about the inappropriate uses of tests warn that if tests are used to bestow rewards or impose sanctions, there are several risks: widening the gap in educational opportunities between haves and have-nots, narrowing curricula, centralizing educational decision making, and deprofessionalizing teachers.

Much of the controversy around testing is not about the tests themselves, but about how their results are used. A test whose scores are proven to be reliable, valid, and fair for a certain purpose, such as providing feedback about the overall level of student achievement in a school or district, may be inappropriate when used to determine if individual students in that district should be promoted to the next grade or receive a high school diploma. Such use may be inappropriate because test scores that have serious consequences for individuals must be more reliable, or precise, than test scores that are used to gauge overall performance of a large number of students.

As the consequences of tests become more serious in this country, much of the public controversy about tests centers around how they are being used: Should a single test be used to make life-changing decisions for individuals? Should students with special needs be exempt from certain tests? How much confidence can one have in test results for decisions about the quality of teachers, schools, districts, or state education programs?

In many situations, standardized tests provide the most objective way to compare the performance of a large group of examinees across places and times.

Test scores are one relatively efficient and objective source of information for helping to make decisions about such issues as course placement, grade promotion, graduation, college admissions, and the competency of teacher candidates. Such decisions will be made with or without tests, so proposed alternatives to the use of test scores should be at least equally accurate, efficient, and fair. *Myths and Tradeoffs* (1999) reminds readers that the U.S. K-12 education system is characterized by variety and decentralization. Curricula, grading standards, and course content vary enormously from school to school. Standardized college admissions tests supplement grades and other information by providing a common yardstick for comparing students from diverse schools with different grading standards. For college admissions officers, standardized tests are an efficient source of comparative information about students' ability to do college-level work, for which there is currently no substitute. Similarly, in K-12 education, statewide standardized tests are useful for comparing student achievement across classrooms, schools or districts, or across different times. When combined with other sources of information, these comparisons can help educators and policy makers decide how to target resources.

A test score is an estimate rather than an exact measure of what a person knows and can do.

A typical large-scale test goes through extensive research, development, and pilot testing to make sure it is an accurate measure of examinees' competencies. Yet despite the best research efforts and most up-to-date testing technologies, test scores always include some "measurement error" due to factors unrelated to student learning. For instance, there is measurement error related to the fact that the questions on a test are only a sample of all the knowledge and skills in the subject being tested—there will always be students who would have scored higher if a particular test version had included a different sample of questions that happened to hit on topics they knew well. Other examples of factors that contribute to measurement error are students' lucky guesses, physical condition or state of mind, motivation, and distractions during testing, as well as scoring errors. Therefore, a test score is not a perfect reflection of student achievement or learning.

High-stakes decisions about individuals should not be made on the basis of a single test score.

One common problem is the tendency to use what are single, inexact measures to make very important decisions about individuals, for instance about promoting students to the next grade or awarding them a high school diploma. Testing professionals advise that when making high-stakes decisions it is important to use multiple indicators of a person's competency, which enhances the overall validity (or defensibility) of the decisions based on the measurements. It also affords the test taker different modes of demonstrating performance.

Systems for State Science Assessment (2005) promotes systems of assessments that might include classroom-, district-, and state-level measures that are aligned to the same set of learning standards and provide different sources of information about student performance. There are many ways that multiple indicators can be considered. For example, students who perform poorly on a statewide high school exit exam might be able to demonstrate their competence by submitting a portfolio of their classroom work.

Tests should not be used for high-stakes decisions if test takers have not had an opportunity to learn the material on which they will be tested.

High Stakes (1999) concludes that tests should be used for important decisions about individual students only after implementing changes in teaching and curriculum that ensure that students have been taught the material on which they will be tested.

Many states, for instance, plan a gap between introducing a new high school exit exam and actually having it count toward graduation, with the expectation that during the phase-in period schools will achieve the necessary alignment among the tests, curriculum, and instruction. Such alignment helps to make a testing program educationally sound and legally defensible. But other decision makers may see attaching high stakes to individual student test scores as a way of leading curricular reform, not recognizing the danger that such uses of tests may lack the necessary degree of alignment between the material being tested and the material being taught.

States, districts, and schools should aim to maximize the participation of English-language learners and students with disabilities in large-scale tests.

Standards-based education reforms seek to apply the same high standards and assessments to all students, including students with disabilities and English-language learners. At the same time, legal requirements stress the individualization of instruction for students with special needs. To what extent can the goals of common standards and assessment and individualized education be reconciled? *Educating One and All* (1997) and *Keeping Score for All* (2004) address this problem. Ideally, students with disabilities and English-language learners should be tested in a manner that provides appropriate accommodation for their special needs while maintaining the validity of the test results. Accommodations are changes in the testing situation that make it possible for students with special needs to participate meaningfully in a test. They are intended to make a student's disability or language status less of a factor in measuring academic performance. For instance, a student with a reading disability might be read a math test aloud. However, if the test assesses the ability to do advanced word problems, then reading the test to the examinee may be inappropriate because it actually changes what is being measured.

Determining which accommodations are appropriate for whom and under which circumstances is difficult—there is limited research about how different types of accommodations affect the validity of test scores. The two objectives of maximizing participation and ensuring the validity of test results for special needs students are sometimes in tension. Policy makers should bear in mind that students with disabilities and English-language learners are particularly vulnerable to potential negative consequences when high-stakes decisions are based on tests.

Teachers need professional development that helps them better understand core principles of assessment and how to apply these to their regular instruction and testing.

Currently, educational policy makers assign much greater value and credibility to external, large-scale assessments of individuals and programs than to classroom assessments that are designed to assist learning. But classroom assessment is critically important for learning. *Assessment in Support of Instruction and Learning* (2003) shows that children learn more if instruction and assessment are integrally related. Teachers can

maximize learning when they provide students, as they are learning, with feedback about particular qualities of their work and about what they can do to improve. More research, development, and training investment must be shifted toward effective use of assessment in the classroom, where teaching and learning occur.

DESIGN

Tests can be a valuable tool and are used for a variety of purposes in education, such as assessing overall student achievement in a school, assisting with the diagnosis of individual learning difficulties, deciding whom to admit to an institution of higher learning or professional school, or certifying that a person is qualified to teach. A single assessment will not be appropriate for all of these uses; the design of a test is largely guided by the purpose it is intended to serve. How does one design a test to ensure that it measures what it is intended to measure and that the conclusions drawn from the test results are justified?

In the design of tests, form must follow function.

It is important for policy makers and test developers to first consider: What is the purpose of this test? What sort of information do we want to draw from the test results? For what decisions will the results be used? In order for a test to be designed well, its intended uses and the kinds of results that will be reported must be carefully considered and articulated, and they must govern the design process. Policy makers should be wary of not being able to articulate a clear purpose for a test or of using one test for many different purposes. In general, the more purposes a single test aims to serve, the more each purpose will be compromised.

The National Assessment of Educational Progress (NAEP) offers an example of the tradeoffs that inevitably arise when designing an assessment. NAEP was intended to survey the knowledge of students across the nation with respect to a broad range of content and skills. *Grading the Nation's Report Card* (1999) and *NAEP Reporting Practices* (2001), the design selected to achieve this is based on sampling—not all students in the country take NAEP, and different students take different questions—enabling NAEP to administer nationwide assessments with hundreds of different tasks. Some have argued that NAEP should test every student and provide student-level results, but to do so would require giving all students equivalent test forms and so limit the breadth of what NAEP could cover. Thus, NAEP's design is specifically suited for giving a snapshot of student achievement across the country, but is not appropriate for measuring student achievement at the individual level.

The design process must ensure that test score interpretations are valid.

Everyone who deals with tests has heard that "validity" is important. In testing, validity refers to the degree to which judgments about students, based on their scores on a particular test, are defensible. For instance, a college admissions officer needs to know that there is indeed a relationship between scores on a college admissions test and performance in college course work. If students who score well on a college entrance exam generally do well in the first year of college, then that is one piece of evidence that the test is valid for college admissions purposes. If the test is a poor predictor of performance in college, its validity for making admissions decisions would be questioned. Similarly, a test used for grade promotion is valid for that purpose only if it clearly assesses fundamental skills and knowledge that virtually everyone would agree are reasonable expectations at that grade level and if students had the opportunity to learn the material that is being tested.

As described in *High Stakes* (1999), there are many different forms of evidence that test developers should collect to determine if a test interpretation is valid—validity is a matter of degree. *High Stakes* also emphasizes a principle that is often neglected—that what needs to be validated is not the test in general, but rather each inference (or judgment) that is made from the test scores and each specific use to which the test scores are put. Although there is a natural tendency to use existing tests for new and different purposes, each new purpose must be validated in its own right.

The design process must ensure that the test results are reliable and fair.

"Reliability" refers to the consistency or reproducibility of a test's results. For instance, a test is highly reliable if a student taking it on two different occasions (with no learning in between) earns essentially the same score or if a person who takes different versions of the same test earns nearly the same score each time. For tests that are scored by hand, such as essays, another important aspect of reliability is the degree to which different raters assign the same score to a student's response.

Fairness in testing encompasses a broad range of issues, including absence of bias in test questions, equitable treatment of all examinees in the assessment process, and opportunity to learn the material being tested. Fairness also refers to comparable validity—if the scores from a test underestimate or overestimate the competencies of

members of any group, the test is unfair. A test is also unfair if it measures different sets of skills for different groups. An example would be a science test that assesses science knowledge for native English speakers, but because of unnecessarily difficult words, assesses both science and language proficiency for English-language learners.

High Stakes (1999) emphasizes that when a test has serious consequences for individuals, it is especially important to have evidence that the test scores are valid, reliable, and fair for a particular use, so that it can stand up to public and legal scrutiny. These qualities of a test cannot be addressed as an afterthought once the test has been developed, administered, and used; they must be confronted from the earliest stages of design.

Testing professionals should consider the relationships among cognition, observation, and interpretation–the "assessment triangle"–when evaluating the soundness of current educational tests or designing new ones.

Knowing What Students Know (2001) lays out the framework in which every educational test rests on three pillars: cognition, observation, and interpretation. These elements are represented as a triangle because each is connected to and dependent upon the other two.

Cognition refers to the process of how students learn and develop competence in a given subject. For example, how do young children gain understanding of numbers? How do they learn about why the number 5 "is greater than" the number 4? How do they mentally represent subtracting one number from another, or counting by 5s or 10s? Research has shown how young children progress from understanding the central concepts of more and less, to counting objects, to being able to do simple addition and subtraction problems in their heads, and so on. Research has also illuminated differences in how novices and experts organize, categorize, and interrelate bits of knowledge in certain subject areas. Research-based descriptions of how children develop understanding of particular subject matter should be the cornerstone of test design. Tests should be designed to differentiate between levels of expertise and to reveal incomplete understandings or misconceptions that students have.

Observation refers to the test questions and tasks that are used to collect evidence about what students know and can do. Observations might be gathered by having students answer multiple-choice questions, write an essay, perform a piece of music, or conduct a science experiment. The choice of observations is not arbitrary; they must be carefully designed to provide evidence that is linked to the known processes of cognition.

Interpretation refers to methods used to analyze the evidence gathered through observations and to combine the information into a score. In large-scale testing, the interpretation of the evidence is usually made using statistical methods, but in classroom testing, the interpretation is often a judgment made by the teacher. The inter-

pretation method must fit with the cognition and observation elements of an assessment: for instance, the cognitive underpinnings of an assessment will guide the selection of an interpretation method by suggesting the most important knowledge and skills that should be highlighted in the test scores.

Test design should begin by making these three elements explicit and ensuring they are compatible. Efforts should be made to be sure that tests reflect recent advances in understanding of learning and how it can be measured.

Advances in the cognitive sciences and measurement offer opportunities to develop educational assessments that better support learning.

Researchers know a great deal about how students learn, and also about how to *measure* what students learn. Several decades of research in the cognitive sciences have improved understanding of how children develop understanding, how people reason, which thinking processes are associated with competent performance, and how knowledge is shaped by social context. At the same time, there have been significant developments in ways to measure student performance using a wide range of statistical methods, and computers are removing many of the constraints that have previously limited assessment practices. *Knowing What Students Know* (2001) describes these innovations and concludes that although most of them have only been tried out on a small scale, they hold promise for a future generation of educational assessments that better inform and support learning.

CONSEQUENCES

Testing can lead to positive consequences when it relates to clear educational purposes and goals as well as what actually happens in the classroom. However, testing programs can have unintended or negative consequences, particularly when some students have not had the opportunity to learn the material being tested or when students have disabilities or are just learning English. For such students, the consequences of high-stakes tests can be quite severe. How can policy makers hold all students to high standards, yet also be fair and equitable to those with fewer opportunities or special needs?

The people who design and mandate tests must be constantly vigilant about equity concerns, including opportunity to learn, cultural bias, or adverse impact.

Sometimes tests are implemented in order to enhance equity by making standards uniform and transparent and by applying the same standards of evaluation to all test takers. However, there may be inequities, such as when some test takers have not had the opportunity to learn the material or when a test incorporates cultural content that disadvantages some test takers. *High Stakes* (1999) emphasizes that it is important to understand that test use may have negative consequences for individual students even while serving important social or educational policy purposes. The development of a testing policy should be sensitive to the balance among the individual and collective costs and benefits of various test uses.

In the absence of effective services for low-performing students, better tests will not lead to better educational outcomes.

Tests, alone, will not help students who are falling behind in the classroom. For instance, as discussed in *High Stakes* (1999), research shows that students are typically hurt by being held back and repeating the same grade in the absence of effective instructional services. Testing can best improve student learning if it is tied directly to efforts to build the capacity of teachers and administrators to improve instruction.

Test results may be invalidated by teaching narrowly to a particular test.

As tests take on greater importance in the United States, "teaching to the test" is becoming more widespread and problematic. People often forget that a test only assesses a sample of students' knowledge and skills in a particular subject. Test results may be invalidated by teaching so narrowly to a particular test that scores are raised without actually improving the broader set of academic skills that the test is intended to measure.

Many people are confused or disagree about what exactly constitutes "bad" teaching to the test. Narrow teaching to the test might include drilling students on practice questions or focusing instruction on the limited subset of skills and knowledge that are most likely to show up on the test. These practices are technically permissible, and might even be appropriate to a limited degree, but they are not likely to provide instruction that helps students understand the material in a way that generalizes to the broader subject domain. In addition, there will be important parts of the curriculum that are not tested and so are neglected. When teachers teach directly to the specific test questions—for instance, ones that have been publicly released from past years' tests—student's test scores are likely to give an inflated picture of students' understanding of the broader subject.

The more ideal situation, as laid out in *Testing, Teaching, and Learning* (1999), requires that states have in place a system of well-designed and aligned standards, curriculum, and assessments, so that teachers have a clear and consistent set of standards to teach toward that are not limited to the content of a previous year's test. In addition, all students should receive sufficient preparation in test-taking skills so that their performance will not be adversely affected by unfamiliarity with a test's format or by ignorance of effective test-taking strategies.

New testing programs should build in an evaluation component.

Testing programs should be evaluated to see if they are achieving their stated purpose. As part of the evaluation, consequences—both positive or intended and negative or unintended—should be carefully monitored and weighed. Newly developed tests, a ubiquitous feature of educational policy today, need to be studied for their impact on particular groups of test takers and their effects on curriculum and instruction. The NRC has conducted several evaluations of national assessment programs, including *Grading the Nation's Report Card* (1999).

PUBLIC UNDERSTANDING

Although there is general public support for tests, there is also significant opposition. Some "anti-testing" groups are quite vocal and have many criticisms of how tests are designed and used; these criticisms deserve consideration and discussion. Some of the opposition can be addressed through adequate public communication by educational leaders as to the purpose of a test, what students are expected to know and be able to do, how the test is administered and scored, how the scores are used, and their consequences.

Test developers and policy makers should clearly explain to the public the purpose for a test and the meaning of different levels of test performance.

In an effort to improve public understanding of test scores, many test findings are now reported using achievement levels such as "basic," "proficient," or "advanced." Decisions about what gets tested and what constitutes "proficient" or "basic" performance on a particular test are the result of lengthy deliberations among educators, policy makers, and test makers. As described in *Measuring Literacy: Performance Levels for Adults* (2005), there are established, systematic methods for guiding the judgment process so that the test results are valid, reliable, and meaningful. However, nonexperts often do not understand that the process used to set achievement levels rests largely on the informed judgment of experts; instead, people often assume the levels reflect some absolute "truth" about what constitutes proficient or advanced performance. Proper reporting of test results requires that the meaning of achievement levels should be communicated clearly. *Grading the Nation's Report Card* (1999) examines the achievement levels set for the National Assessment of Educational Progress.

Not only do large-scale tests provide means for reporting on student achievement, but they also convey powerful messages about the kinds of learning valued by society. Policy makers and educators need to communicate among themselves, and to the public, the kinds of thinking and learning they want to encourage in students. Content standards and sample test questions accompanied by student responses representing different levels of competence should be shared with the media, parents, and students. In this way, tests can foster dialogue about the larger issue of what students should know and be able to do.

When test results are reported to students, teachers, and the public, the limitations of the test should be explained clearly to a lay audience.

Test designers and policy makers should explain to test consumers that no test is a perfect measure of what an examinee knows and can do, that multiple indicators lead to more valid decisions, and that the questions on a given test are only a sample from the larger domain of knowledge and skills that students are expected to learn. People affected by tests, including parents, should receive information that explains, in a clear way, such concepts as measurement error, significance of score differences, the probability of misclassification, and, when applicable, how well the test predicts future performance. Such information will help people better understand what the test results really mean.

CONCLUSION

Tests are one objective and efficient way to measure what people know and can do, and they can help make comparisons across large groups of people. However, test scores are not perfect measures: they should be considered with other sources of information when making important decisions about individuals.

FOR FURTHER READING

The testing principles outlined in this booklet are described more fully in the NRC's reports, listed below, which can be purchased and downloaded through the National Academies Press website (http://www.nap.edu). More information about the Board on Testing and Assessment can be found at http://www7.nationalacademies.org/bota.

Assessment in Support of Instruction and Learning: Bridging the Gap Between Large-Scale and Classroom Assessment (2003) As large-scale standardized testing is increased, what can be done to better integrate it with the classroom assessments that teachers use every day? The report explores the strategies used in about a dozen programs that seek to bridge the gap between large-scale and classroom assessment.

Educating One and All: Students with Disabilities and Standards-Based Reform (1997) Standards-based education reforms seek to apply the same high standards and assessments to all students, including students with disabilities. At the same time, legal frameworks for students with disabilities stress the individualization of instruction. To what extent can the goals of common standards and assessment and individualized education be reconciled?

Evaluation of the Voluntary National Tests: Year 1 and Year 2 (1999) In 1997 President Clinton announced a federal initiative to develop national tests of 4th grade reading and 8th grade mathematics. The tests would be voluntary because the federal government would prepare but not require them. Congress called on the NRC to evaluate various aspects of the test development process as it was occurring.

Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress (1999) The National Assessment of Educational Progress (NAEP) has provided data about what American students know and can do for over 30 years. How could NAEP be improved so that it provides more useful information about student achievement?

High Stakes: Testing for Tracking, Promotion, and Graduation (1999) The use of tests to make decisions with important consequences for individual students is growing. Three such decisions involve tracking, grade promotion, and granting a high school diploma. Under what conditions is it appropriate and fair to use tests that carry such high stakes for individual students?

Keeping Score for All: The Effects of Inclusion and Accommodation Policies on Large-Scale Educational Assessment (2004) As efforts are made to include more students with disabilities and English-language learners in educational assessments, leaders are faced with tough decisions about which students should be included and accommodated in testing. The report synthesizes research findings about the effects of accommodations on test performance and reviews current procedures that large-scale testing programs use to make inclusion and accommodation decisions, with a special focus on the implications for the National Assessment of Educational Progress (NAEP).

Knowing What Students Know: The Science and Design of Educational Assessment (2001) Advances in the cognitive sciences have increased understanding of those aspects of learning that are most important to assess. At the same time, advances in measurement and technology permit the collection and interpretation of more complex information about student performance. How can these advances be used to develop new kinds of educational assessments that better support learning?

Measuring Literacy: Performance Levels for Adults (2005) The National Assessment of Adult Literacy is a household survey conducted periodically by the Department of Education to evaluate the literacy skills of a sample of adults in the United States. This report details the process that an NRC committee used to determine the five performance level categories that should be used to characterize adults literacy skills. The report also recommends ways to communicate about adult literacy and improve how it is assessed in the future.

Myths and Tradeoffs: The Role of Tests in Undergraduate Admissions (1999) It is important the college admissions process, especially at elite colleges, be both fair and open. How should test scores be used in the college admissions process?

NAEP Reporting Practices: Investigating District-Level and Market-Basket Reporting (2001) The government sponsors of the National Assessment of Educational Progress (NAEP) have been taking a critical look at their procedures for reporting NAEP results with an eye toward improving the usefulness and interpretability of the results. They asked the NRC to convene a committee to examine the ways in which NAEP reports are used and misused by policy makers, educators, the press, and others and to suggest ways that NAEP reporting could be improved.

THE NATIONAL ACADEMIES
Advisers to the Nation on Science, Engineering, and Medicine

The nation turns to the National Academies—National Academy of Sciences, National Academy of Engineering, Institute of Medicine, and National Research Council—for independent, objective advice on issues that affect people's lives worldwide.
www.national-academies.org