# Test Bias

In J. Worell (Ed.), *The Encyclopedia of Gender* (pp. 1129–1140). Academic Press, 2001.

Marcia C. Linn
4523 Tolman Hall #1670
Graduate School of Education
University of California at Berkeley
Berkeley, CA 94720-1670
e-mail: mclinn@socrates.berkeley.edu
fax: (510) 643-0520

Cathy Kessel
Tolman Hall #1670
Education in Mathematics, Science, and Technology
University of California at Berkeley
Berkeley, CA 94720-1670
e-mail: kessel@soe.berkeley.edu

This material is based upon research supported by the National Science Foundation under grant REC 98-73160 and REC 98-05420. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

The authors appreciate the help and encouragement of the Science Controversies On-line: Partnerships in Education (SCOPE) and Web-based Integrated Science Environment (WISE) project research groups. Special thanks are due to Jacquie Madhok and Ann Shannon for extensive, detailed review.

Preparation of this manuscript was made possible with help from Lisa Safley, David Crowell, and Lisa Bigelow.

# GLOSSARY

Affirmative action: Programs initiated to remedy past discrimination including discrimination based on use of biased tests. These programs include diligence in selection decisions, provision of specialized educational opportunities, and advice or coaching to individuals from groups that have suffered discrimination in the past. The latter are often social or cultural, e.g., women, girls, and racial, ethnic, or low-income groups.

Criterion-referenced test: A test in which success or failure is determined according to standards concerning the content of the test rather than with reference to the scores of a given population as do norm-referenced tests. In contrast with norm-referenced tests, the design and scoring of criterion-referenced tests allow the possibility that every test taker receives the highest possible score.

Educationally valid: An educationally valid measure is: a) sensitive to the program for which it is used, b) assesses test takers' abilities to monitor their own performances, c) promotes lifelong learning. Educationally valid measures draw on a mixture of performances. Their validity is determined by corroborating longitudinal studies rather than internal consistency.

High-stakes assessment: We refer to assessments as high stakes if they are traditionally used for decisions that have major consequences for the test-takers or for others involved with the test-takers such as teachers, principals, or superintendents. Assessments are only high stakes if they result in important consequences for the participants. For example, university entrance examinations are often considered high-stakes assessments, but some universities do not use scores from these examinations in admissions decisions.

Norm-referenced test: A test that is normed on representative test takers and periodically recalibrated in order to maintain a normal distribution of scores, in which half of the test takers receive scores that are below average.

Question or item context: An item or question may have a setting intended to be independent of the ability to be assessed. The “cover stories” given by word problems in mathematics are a common example. These settings are known as item or question contexts.

Social context: Most testing activities and interpretations occur in a context involving other people. Social context is important because it often creates expectations and influences reactions in the test-takers, scorers, and score users.

Standardized test: A test that is administered under established conditions, scored consistently, and shown to yield reliable performance across time. Often but not always these tests use a multiple choice format.

Test: A collection of questions or tasks designed to measure abilities and skills. Testing professionals also include in this definition the stipulation that responses be scored and evaluated in a standardized way. Oral examinations are one example of an evaluative device that often does not satisfy the latter stipulation.

Stereotype threat: Groups are sometimes stereotyped as lacking or deficient in particular abilities. A stereotype threat situation, for example, the announcement that an exam tends to show gender or race differences in average scores, triggers consciousness of that stereotype.

Test bias: Tests are biased when irrelevant or systematic factors skew performance or scores for all or some of the test-takers. Bias can result from numerous factors including the social context, question or item context, and setting of test administration.

University entrance examinations: Tests such as the GRE, SAT, and ACT that are used in admissions decisions for undergraduate and graduate programs.

Reliable measure:  A measure is reliable if it consistently yields the same score for the same individual independent of variation in time of day, context, or other irrelevant factors.

# Introduction

"Don't ask me anything, I can't do science."

"Don't believe my test score. I can't do tests. Look at my grades, my recommendations, or my science fair project."

"How could I succeed on that test—it didn't measure anything that was taught in our school."

"Our school spends more on testing than on curriculum. Instead of buying materials for science experiments, we buy more tests."

"All the poor children failed—the test measures income, not achievement at our school."

"I can't teach science with projects any more—there are too many topics to cover for the test."

"Reading tests determine the school budget. I am not encouraged to teach science."

"The tests give you the answers—all you have to do is select among the options; our students think memorizing is the best way to learn."

Individuals often question their own test scores, regularly explaining that a test does not accurately capture their true abilities, that their performance was unfairly influenced by external conditions, or that the test itself is a poor indicator of potential. Few argue that a test has underestimated their performance. For example, a growing proportion of individuals have been labeled learning disabled, dyslexic, or attention impaired to exempt them from typical testing constraints. At the same time, more and more people call for high-stakes tests to determine everything from individual student promotion, to teacher salaries, to statewide school budgets. Although teachers complain that tests are frequently insensitive to instructional innovation and poorly aligned with curriculum, policy makers rightfully argue that performance is too low and individuals are poorly prepared for today's workplace. Paradoxically, many complain that tests do not accurately measure their ability, saying "I'm a poor test taker" and "This test doesn't show what I can do," while still endorsing the use of tests for practically everyone else.

Historically, ability tests have added a dimension of objectivity to decisions that were often made behind closed doors and based on favoritism, patronage, or stereotype. Tests have played a positive role in increasing access of women to medical school, law school, and higher education . The United States was among the first countries to promote extensive public education. Standardized testing began in an attempt to monitor the quality of this education. Tests were also used to sort children by "aptitude" (measured by I. Q. tests) and "ability" (measured by achievement tests).

Increasing reliance on tests raises very complex issues about the systemic, organic, convoluted institution of education. The effectiveness and bias of a test depends[1] on its role in the educational system. Performance on tests reflects at least the quality of instruction, the opportunity to learn, the validity of the test, and the context of assessment. Much rhetoric touts a "deficit model" to explain why one group, such as men or boys, outperforms another, such as women or girls, on any assessment. Followers of this model argue that unsuccessful groups lack a particular ability, skill, or capability that should either be remediated or accepted. An alternative explanation looks more carefully at tests and performance, stresses that there are "multiple paths" to success in a given area, checks to see if the instruction can improve performance of all students, and seeks to ensure that all students have the opportunity to learn.

*Educational validity*

Rather than focusing on tests irrespective of use, we consider what we call educationally valid uses of tests. Evaluating the bias of a test, or its educational validity, requires examining all the factors contributing to student success or to decisions about individuals and groups, rather than looking solely at the test. Thus, tests are only as valid as the uses to which they are put. Using a measure of height to determine presidential candidate success, for example, has historically proven to be accurate, but has no validity in terms of either instruction or prediction of success as a president. Tests that inadvertently draw, for example, on ability to compute baseball statistics might be valid for individuals in a culture where everybody follows baseball, but ineffective when individuals who lack this

opportunity are included among test-takers. Even the aspects of learning that tests measure may vary by social or cultural context. For example, Ann Gallagher studied students who received high SAT scores. She found that females often used more time-consuming school-taught techniques to answer SAT items while males tended to use short-cuts.

Today, increased enthusiasm for high-stakes tests in the United States has heightened the importance of making the uses of these tests educationally valid and of ensuring that programs designed to improve performance can succeed for the diverse learners in our schools. The eagerness of policy makers to use tests for high-stakes decisions including promotion, tracking, and graduation intensifies the importance of careful research and analysis of the impact of tests. Recent reports from the National Research Council on high-stakes testing reveal the complexities and difficulties of these issues. A particular concern has been the instructional sensitivity of tests and the effectiveness of programs for ameliorating difficulties revealed by tests. For example, longitudinal studies suggest that students from lower income families compared to students from higher income families are more negatively impacted by elementary school teach-to-test methods and more advantaged by developmentally appropriate programs.

Ideally, all tests would be concurrently designed with the programs they evaluate to ensure alignment between instruction and assessment. This means researching the alignment to ensure that tests are sensitive to variations in instruction and that students gain capabilities that will serve them well in the future. Programs that reward

individuals, teachers, or schools for successful performance on tests need to demonstrate that improved instruction leads to improved performance both immediately and cumulatively. All too often, students are assessed on performances they have no opportunity to acquire or on dimensions that poorly predict future success.

The costs in lost opportunities and unintended consequences of denying promotion, censoring teachers, and discontinuing programs must be carefully analyzed. How can we decide whether students at a low-performing school should drill on basics or read rich, culturally sensitive accounts of individuals who have overcome unbelievable odds to succeed? What benefits arise from retaining middle school students in classes that have already failed them, especially when they drop out of school after ninth grade instead of tenth grade? Can we be sure that students assigned to remedial programs based on test scores will succeed, or conclude they cannot understand the topic—or cannot learn at all?

We must carefully weigh the strengths of the tests that we use for instructional, personnel, and admissions decisions, and look at the tradeoffs between individual success and group progress. Many acknowledge this dilemma by calling for the use of multiple indicators, yet this is only a first step towards solving an extremely complex problem. We certainly need multiple indicators, but we also need to develop wisdom about their use.

For example, multiple indicators can be misused as Barbara Bergman, in her book *In Defense of Affirmative Action,* has pointed out. Selection committees get around affirmative action constraints by utilizing different indicators depending on which

candidate they wish to remove from a pool. Individual candidates might be rejected piecemeal by the use of a particular criterion for each. However, the successful candidate could conceivably not satisfy all of these criteria.

*Test Bias*

Test bias results when indicators intended to improve selection of individuals or evaluation of programs are perturbed by irrelevant or inappropriate factors that result in flawed decisions. Such factors may occur in the design, administration, or use of a test. For example, the design of admissions tests may inadvertently advantage individuals who have grown up in a particular culture and be biased against members of other cultures. One famous example concerns a vocabulary item from the Scholastic Aptitude Test that required students to know the meaning of "regatta"—a term more likely to be encountered among upper-middle-class European-Americans than other groups.

When tests determine the fate of programs intended to give students a head start or ameliorate injustices, bias in measurement may result in denial of services to those most in need of the treatment. For example, reading programs that teach students to use contextual cues cannot be properly evaluated by tests with items about the meanings of isolated words.

A poorly designed test may also fail to detect undesirable teaching methods. The "key word" approach for solving elementary mathematics word problems illustrates the point. In this instructional approach, students learn to translate words into operations. These

students learn, for example, to perform subtraction when they encounter the word “left” in the problem statement. Such strategies may lead to short-term success on problems written following the rules but fail when students compute the campaign expenses for "left-wing" causes. We invite the reader to create items that detect the flaws in the key word approach to solving word problems.

Tests may give factors indicative of success too much weight in decisions. For example, rapid computation may be a valuable mathematical skill but college and graduate school admissions tests that reward only speed and accuracy in solving 25 to 35 problems in 30 minutes may neglect the sustained problem solving essential for long-term success in college and in careers. Consistent with this, studies have found that for a given college major, SAT scores tend to underpredict women's grades relative to those of men.

Tests may also mislead test-takers about the capabilities and skills necessary for performing in settings like graduate school and the workplace. Tests can reinforce beliefs that success is dependent on a single inherent ability rather than comprised of an assortment of complex capabilities that can be learned as they are required. Complex tasks such as researching and writing a dissertation, administering a school, or running a medical practice require a broad range of skills and the ability to learn new technologies, policies, and whole fields such as electronic commerce. Recent research by Robert Sternberg and Howard Gardner on the constellation of abilities held by successful individuals underscores this view.

Often biased tests reinforce stereotypes about which groups can succeed in a given endeavor. College admissions tests in mathematics resonate with material in the popular press. Numerous slice-of-life advertisements, situation comedies, and even children's programming depict women as unable to balance checkbooks or do arithmetic. News accounts often reinforce these stereotypes and have even convinced some young women to avoid mathematics courses, in spite of the well-established finding that, on average, women earn higher grades in all mathematics courses. Blatant, subtle, and ubiquitous cultural expectations can have widespread impact on all—including test-takers and test users.

Assessments that rely on a single indicator may inadvertently be biased against individuals who are capable of counteracting weaknesses and utilizing their array of talents effectively for achieving a goal. This is another reason for judicious use of a repertoire of indicators for educational and workplace decisions. Combining mathematics scores and mathematics grades, for example, helps to reduce bias in selection decisions. Understanding the biases of individual indicators and having a sense of the multiple paths that lead to success will help decision makers create repertoires of measures that at the same time encourage lifelong learning and more accurately predict success.

In summary, test bias refers to the systematic over- or under-prediction of success in a given educational context. Bias arises when tests measure irrelevant, counterproductive, or unsystematic factors. The costs of test bias to individuals, programs, populations, and

society as a whole can be substantial, devastating, and even life threatening (for example, consider the selection of medical school candidates or prospective engineers). Determining bias requires our understanding of the goals of education and the programs likely to promote success. We turn to a framework for educational validity and then analyze test design from this framework.

## A framework for educational validity

In a given instructional context, tests are educational valid are sensitive to consistent with the goals of instruction, assess learner abilities to monitor their own performance, and promote lifelong learning. These tests draw on a mix of performances to tap the multiple paths to success, rather than relying on only the ideal or typical path. Their validity rests on longitudinal studies rather than solely on internal consistency.

Lifelong learning is of particular importance because today's learners face an uncertain future in which they are likely to change jobs and even fields regularly. The demands of the workplace ensure that individuals will need to learn how to use new technological tools, like personal computers, master new communication skills such as video conferencing, understand new, complex issues such as genetic engineering or space exploration, and flexibly adapt to as yet unanticipated conditions.

Cognitive scientists have begun to characterize lifelong learning and to question prior assumptions that lifelong learning capabilities were rooted in the learning of subjects such as Latin or computer science. Research suggests that preparing individuals to master a new complex topic in depth requires that they experience mastering similar topics in depth, learn to conceptualize projects and identify useful resources, develop the ability to monitor and evaluate their own progress, and become adept at designing opportunities to learn from others when their progress falters.

Educational programs that promote lifelong learning engage students in sustained reasoning about a topic from their discipline while at the same time providing opportunities to learn contemporary skills such as searching for information on the Internet or negotiating with a healthcare provider. Individuals also need understanding of fundamental ideas that permeate our culture such as electronic communication, social justice, and environmental stewardship. Research shows that students gain this kind of understanding from courses and advanced programs that require extended projects, iterative refinement of solutions to complex problems, and negotiation about these solutions with others. Courses with projects, analysis of cases, and research investigations prepare students for handling complex problems in future programs and the workplace.

Designing tests to measure this kind of understanding means addressing a number of complex issues. Too often tests encourage superficial understanding rather than linked and connected ideas combined with the ability to guide one's own learning suggested by

the goal of becoming a lifelong learner. Tests serve as a model for course activities and have the potential of reinforcing instruction. We need lifelong-learning assessments aligned with instruction that promotes lifelong learning. Recall and vocabulary tests not only fail to encourage lifelong learning, they frustrate teachers and students.

Linn and Hsi, in their book *Computers, Teachers, Peers—Science Learning Partners*, use case studies, classroom tests, and class projects to illustrate instruction for lifelong learning. Students in the semester-long Computer as Learning Partner course continue to build up their understanding of thermodynamics as they progress from middle school to high school. No such evidence of lifelong learning was shown by a comparison group in the traditional vocabulary-driven curriculum.

The performance of students from the Computer as Learning Partner curriculum in twelfth grade demonstrates the importance of assessments that stress lifelong learning rather than requiring superficial multiple choice answers. In interviews asking about complex problems, students who initially seem confused talk about the various alternative interpretations of a problem, compare several potential responses, and eventually create a coherent explanation based on multiple ideas. For example, Linn and Hsi describe four students who attempt to explain a conundrum: Why do metal objects feel colder than wooden objects at room temperature but, when tested, measure the same temperature? In tenth grade one of these students (see page 263) relies on the result of the test and offers no explanation, saying they measure the same temperature. In twelfth

grade this student elaborates to explain the conundrum, gives a somewhat confused answer, but makes some promising connections that bode well for future learning.

Using a traditional multiple choice test could mask this progress and might also reinforce superficial thinking. An essay asking for an explanation would award the tenth grade answer minimal credit and give the twelfth grade answer a higher score because the student combined ideas and sought to resolve the conundrum. The multiple-choice approach could discourage lifelong learning and reinforce student reliance on memorization, while the essay question has the potential of encouraging coherent understanding. The essay question is more sensitive to instruction that promotes lifelong learning, while multiple choice exams have a potential bias against lifelong learning.

Tests that seem to require superficial understanding of topics have the potential of reinforcing instructional programs emphasizing superficial knowledge. Teaching superficial understanding is clearly more straightforward than attempting to help students understand complex, nuanced, and uncertain phenomena. Many science texts, for example, pay no attention to current scientific controversies and instead emphasize "established" findings. Students respond to these texts by memorizing information about science. Under good instructional conditions, students remember the material they memorized at least until they take the test. However, research shows that this form of instruction results in rapid forgetting. As a result, American students perform relatively poorly on national assessments and particularly poorly in areas where cumulative understanding is crucial.

The Third International Mathematics and Science Study results for students in high school physics and calculus, for example, shocked leaders in science and mathematics by revealing that American students were performing in the lower third of countries when only elite students were compared. These results suggest that somehow the U.S. educational system is failing to instill the cumulative understanding of these topics characteristic of instruction in other countries. Laudably, the TIMSS test included complex problem solving and short essays as well as multiple choice items. U.S. students performed poorly on all aspects of these topics, but were particularly disadvantaged in the more complex accomplishments.

*Bias in Design*

Test designers often make seemingly innocuous decisions that result in biased scores for some test-takers. For example, test designers often ask respondents to provide demographic information prior to starting the test. The most intriguing finding concerning the impact this may have comes from the psychological experiments of Claude Steele and his collaborators. Steele's research program demonstrates that simply reminding students that they belong to a cultural group stereotyped as unlikely to succeed on a high-stakes speeded, standardized assessment like the Graduate Record Examination can dampen their performance on tests.

Steele has identified "stereotype threat" to explain why individuals aware that their cultural group is often less successful in a given endeavor are likely to perform less well.

Several mechanisms have been put forth to account for this situation. Under one scenario, individuals reminded of stereotype threat become more anxious and perform less well due to debilitating anxiety. Under another model, individuals informed of stereotype threat lose motivation and try less hard because they assume the outcome of their effort. A third possible mechanism concerns playing it safe or taking fewer risks as a result of stereotype threat. Individuals choosing a safe approach might revert back to more algorithmic and less creative problem solutions or check their work extra times, thus perhaps performing less well on tests where efficiency and speed are rewarded.

Decisions about test format may advantage students with more experience, for example in taking timed tests or using scoring sheets with bubbles. Redesign of the Graduate Record Examination as a computer-adaptive test revealed previously unstudied aspects of the test.

The computer adaptive version of the GRE revealed that many respondents were unable to finish the test, and especially the analysis section, in the allotted amount of time. The computer administration allowed collection of response latencies by item and showed that for some items responses were given in less time than necessary to read the item. Previous work masked this finding because the paper and pencil version does not penalize test-takers for guessing. Converting to a computer-adaptive format required reducing or eliminating guessing since the response to each item determines the difficulty of the next item administered. This meant that students responded to fewer items, increasing the possibility that the test would yield an unreliable score. In addition, the

computer format was unfamiliar to some test-takers, introducing another potential influence on performance.

In one well-publicized case an individual received very different scores on the two versions of the test. Amy Cuddy took the computer version of the Graduate Record Examination in October 1998 and was surprised to receive scores considerably lower than her practice exam scores. She took the test again in paper-and-pencil format and her scores increased: on the analytic section of the exam, her scores for computer and paper-and-pencil formats were 300 and 690. Eventually, her lower scores were invalidated.

Research in the United States, Australia and Europe shows that question format influences who succeeds. On average, women are more successful than men on essay questions and in large projects, and men are more successful than women on multiple choice questions and, often, in oral exams. Caroline Gipps and Patricia Murphy demonstrate that British mathematics performance varies by test format. For example, females out-perform males on advanced mathematics tests requiring projects but not on multiple choice assessments.

Question context design decisions have been shown to advantage groups familiar with the content. Males tend to be advantaged by questions involving contexts typed as male: for example, in the SAT-V males are advantaged by reading comprehension questions concerning science articles. However, question context may play a more subtle role: for example, a study done by the Educational Testing Service found that females responding

to a GRE-M item were advantaged if the item was set in a business context while males were advantaged if the item was set in a physics context (see Figure 1). Research on the GRE demonstrated that students, in general, have an advantage on reading comprehension questions in their broad major field.

[Here's where the figure could go. Figures A1 and A3 from ETS FAME booklet appendix, p. 9. Note that Figure A1 has a typographical error. The equation that begins *P* = 760 = etc. should be *P* = 760 + etc.]

*Bias in Scoring*

Scoring methods may incorporate biases. For example, when examinations are scored in an unstandardized way, the scorers' biases may prevail. Psychological experiments have established that raters assign different scores to identical performances on written work depending on the gender of the performer or supposed author of the work.

Research suggests that when scorers of complex accomplishments such as essays are aware of the gender of the authors, their scoring decisions are impacted. In general, knowledge that the author is a female results in a lower score than knowledge that the author is a male. This bias arises even when scorers are reading exactly the same response. Because females tend to outperform males on essay questions, this bias suggests that females may perform even more successfully than their scores reveal. This gender bias creates special problems in oral exams. For example, a female graduate student we will call Carole failed her oral qualifying examination because she didn't

answer enough questions in the time allotted. However, Carole was aware that a male student in her department also had not answered "enough" questions on his oral qualifying examination and was given additional time. She appealed the decision, and, after several months, the decision was overturned.

*Bias in score use and interpretation*

Studies of entrance examinations like the SAT and GRE raise several concerns about score use. The SAT was designed for use in undergraduate admissions and validated by showing that it adequately predicts first year college grades. The correlation of SAT score with overall undergraduate grades, however, varies considerably with college major. SAT scores tend to under-predict women's grades relative to those of men. The GRE is designed for use in graduate school admissions and has been validated by showing adequate correlations with first year graduate school grades. The GRE scores of older graduate students, however, tend to under-predict their grades.

Individuals tend to interpret scores according to their preconceptions. Because mathematics in often viewed as a male domain, male success tends to be interpreted as due to ability and female success as due to effort. Researcher Meredith Kimball noted that this process occurred in reports from the Study for Mathematically Precocious Youth (SMPY). Students who volunteer to take the SAT-M in Grades 7 and 8 and receive scores above a certain cut-off point qualify for SMPY. Among qualifiers there are a few girls, but the girls receive substantially higher grades in the SMPY program. Rather than question the selection process, SMPY researchers viewed SAT-M scores as determining

mathematical ability and girls' higher grades as reflecting the girls' better "conduct and demeanor," rather than their mathematical ability.

*Designing the social context*

By defining educational validity in terms of lifelong learning, we emphasize the need for courses to promote substantial reasoning and instructions that provide feedback on progress. For males and females this may mean the social context of instruction may be as important a predictor of success as grades or admissions scores. Research suggests that some contexts are more productive for females than others. Mathematics faculty at colleges where women students predominate report, for example, that they counsel their most talented undergraduates to select from graduate programs where women have succeeded in the past. Graduate programs in mathematics vary dramatically with regard to the number of females that they admit, and even more dramatically with regard to the likelihood that females will complete the mathematics program. Informed mathematics faculty from a number of all-female undergraduate institutions report encouraging their students to attend certain graduate programs, and discouraging them from considering others. They frequently encourage their most talented undergraduates to contact alumnae in graduate school in order to get a sense of the social context for learning in different graduate programs. These counselors report excellent consistency between the program selected and the ultimate success of their graduates, finding this factor to be far more consequential than student performance during the undergraduate years.

Assessing the potential bias in tests associated with education raises special issues for gender equity as well as more general issues where gender equity is an interactive factor. As these examples suggest, attending to gender equity will enable more effective assessment of the educational validity of tests used in consequential decision making.

*Designing for lifelong learning*

Tests are proxies for more complex behaviors and predictions of future complex behaviors. Designing tests to measure lifelong learning that are valid and reliable, and lack unintended negative consequences has proven difficult.

Concurrent design of instruction and assessment can increase alignment between what is taught and what is measured on tests. Today, alignment is frequently neglected. Some states and countries decree frameworks for instruction but leave testing and curriculum design open. In the United States, curriculum design and test design is often disconnected. This lack of linkage between instruction and assessment means that frequently students and teachers are held accountable on tests that have little connection to their textbooks and curriculum materials.

Often the problem is even more severe because the high-stakes tests suggest a form of teaching and learning incompatible with deep understanding and lifelong learning. For example, multiple choice questions, which are inexpensively and easily scored, can take a scattershot approach to curriculum topics in order to have appeal to groups with diverse curricula, for example, different states with different curriculum frameworks. When

teachers look at these tests they may alter instruction to eliminate emphasis on lifelong learning and focus instead on short-term and superficial understanding, including drill on vocabulary. Similarly, students, expecting multiple choice tests, may choose a superficial approach to learning material that seems to be directly represented in the test. Even when test questions require more complex performances, research has shown that instructors can teach to the test in very specific ways, pointing out pitfalls in format and scoring that bypass the understanding the test is intended to measure. For example, in 1988 Alan Schoenfeld studied the classroom of a teacher who was very successful in having students pass the New York Regents Exam. Students focused on memorizing geometric constructions and drawing them accurately, rather than understanding connections between the constructions and proof.

Much recent research to create measures of complex understanding has investigated innovative test formats such as portfolio assessments, performance assessments, group assessments, and projects. Psychometricians analyzing these innovative formats have noted a number of threats to test reliability. Frequently, such formats concern a particular content area and either penalize students unfamiliar with that area or challenge those developing scoring methods to equate performances involving different topics. The problem is further compounded when individual scores are required for group projects, and where conditions of work are rich and varied, meaning that some students may receive more assistance in performing the task than others. Tradeoffs between validity and reliability may result. Many see reliability as the more valuable goal and prefer tests that have high reliability because they fairly reproduce scores from one interval to

another. Others see deep understanding of the topic as the highest goal and prefer tests that measure whether students have gained a kind of understanding necessary for lifelong learning.

This problem of varied item formats is confounded by the underlying factors that contribute to the reliability of multiple choice tests. Researchers have shown that the similarity in performance from one test to another frequently relies far more on knowledge of vocabulary than understanding of the subject matter. Students with a more varied vocabulary may perform well on assessments in mathematics, science, social studies, and English, even though their actual grades and complex performances in these topics vary considerably. Reliance on high-stakes multiple choice tests that are heavily influenced by vocabulary has motivated parents and representatives of dyslexic and learning disabled students to complain that these tests unduly burden slow readers who may have less extensive vocabularies, but are fully capable of performing as well or better than their peers as long as the test measures their ability to solve a complex problem in the domain, rather than their ability to quickly define esoteric vocabulary words in a multiple choice question.

The use of innovative formats and the assessment of complex understanding are further exacerbated by the likelihood that achieving deep understanding and the ability to engage in lifelong learning may occur along a number of different paths. One student may gain understanding by reading extensively, while another may gain it by conducting empirical

tests and reflecting on the results, and still another student may learn by working in a collaborative group.

Designing assessments that honor these diverse paths to success while at the same time measuring the complex kinds of understanding necessary to become a lifelong learner requires research in classrooms. For example, in science learning, some students draw on technology concepts as well as literacy skills to read, interpret, and critique scientific information. Other students may find ways to achieve deep understanding of a topic without the associated technology or communication skills. Test designers need to pinpoint the relevant aspects of performance and tease out the irrelevant influences of question format and disciplinary focus to ensure that individuals who followed diverse paths to success are not penalized.

Many hope that tests, especially high-stakes tests, will drive educational reform. However, designing tests disconnected from opportunity to learn may neglect educationally valid accomplishments in favor of material that is difficult to learn. Biographies of successful entrepreneurs frequently point out that these individuals failed science classes and math classes, dropped out of school, and even today lack understanding of some of the topics that test developers consider essential. Ensuring that tests measure accomplishments we would like citizens to have and we know citizens can achieve is essential for achieving educational validity. Ultimately, tests should encourage students to develop capabilities that will serve them well in the future. Already the crowded curriculum in most disciplines discourages personal reflection and connections

across topics. We need a more frugal curriculum and a more determined reliance on lifelong learning to prepare students for our uncertain future.

*Designing instructionally sensitive tests*

Ideally, concurrent design of assessment and instruction will result in continuous improvement of both the assessment and the instruction. This means that we will design curriculum materials that bend to the needs and demands of teachers and students, but do not break under alternative instructional conditions. It means designing assessments that provide valid and useful information for teachers and students that will allow them to make improvements in the instructional program or their own understanding.

Aligning instruction, assessment, and the social context of learning has the potential of adequately preparing lifelong learners but is difficult and costly in time and materials. Ideally programs of instruction will permit customization by teachers to the context of learning. Thus, if students are studying water quality, an ideal customization would be to tailor instruction to the specific water quality issues in their community. Similarly, if students are studying eighteenth century literature, they might read works by authors from their geographical area that evoke the life and environment characteristic of their community in the past. Designing curriculum materials that are aligned with high-stakes assessments while at the same time, sensitive to differences in student backgrounds and student learning context, remains an active area of research that has potentially great benefit for promoting lifelong learning.

Lifelong learning is best promoted if students have the opportunity to revisit the ideas that they learn in school after they complete their courses. There is never enough time in the curriculum to enable students to learn all the material they need to know. Rather, successful educational programs motivate students to continue to learn about a topic long after they have completed instruction.

Modifying the curriculum to emphasize contemporary issues in mathematics and science and to encourage learners to revisit their ideas has the potential of improving the lifelong learning outcomes of instruction. In order to align such a curriculum with valid tests, we need research programs that emphasize concurrent design of instruction and assessment. We need to identify successful ways to equate responses from students in different geographical areas who have experienced specific customizations of instruction.

In summary, assessment that promotes effective curriculum design and continuous refinement will enable teachers to use results to improve their curriculum and enable students to monitor their own progress. Such instruction will also incorporate generative and effective uses of modern technologies and valid aspects of language literacy. At the same time, these programs and assessments will be sensitive to the varied paths that students take to achieve lifelong learning and that offer mechanisms for assessing students who follow unique paths to success.

*Using test format to promote learning*

Concurrent design of instruction and assessment in the context of a particular school and classroom will enable teachers to contribute as much as possible to the success of their students. Ideally, assessment measures will require students to produce portfolios or engage in projects and thereby reinforce the kind of instruction most likely to lead to lifelong learning.

Some teachers not only teach to the content of a test, but also spend considerable time helping their students become familiar with test format. A large body of research demonstrates that students benefit from practice on tests because they have the opportunity to learn how to respond to new formats. An extreme implication of this finding arises when visitors to elementary school classrooms discover that teachers are devoting up to an hour a day to helping students learn how to "bubble in" responses on answer sheets. Research has clearly demonstrated the benefit of this kind of instruction, but it raises serious problems in regard to tradeoffs between logistic and intellectual activities in classrooms.

Conversely, when tests require essays, students' learning may benefit from practice with the test format. For example, regularly writing reflective essays about their books or investigations has the potential to promote students' understanding of the topic. This instruction can also instill standards of coherence and encourage lifelong interest in connecting ideas.

Several states are currently experimenting with the option for teachers to customize instruction to specific topics and to administer tests that are similarly customized. In Michigan, teachers can select one topic from a list for in-depth specialization. Students spend more time on this topic and the assessment for their performance emphasizes the topic proportionally.

## Recommendations

Designing instruction, assessment, and opportunities for teachers and students to customize their learning so that these activities are aligned has considerable promise for reducing test bias and increasing educational validity. Because education is a complex and intricately connected field, such activities are inevitably complex. To achieve the desired forms of alignment it is clearly essential that teachers, administrators, and innovators work in partnership, respect each other's expertise, and regularly refine educational programs. Building in methodologies for continuous improvement for both assessment and instruction will greatly enhance the effectiveness of new policies. It makes little sense to change tests and assume that students and teachers will suddenly develop the ability to respond to them, just as it makes little sense to assume that individuals can become lifelong learners by neglecting cumulative understanding.

A particular danger of increased reliance on high-stakes tests is that the deficit model of performance will be reified and reinforce stereotypic views of who can succeed in particular fields, endangering opportunities for all learners to achieve the ability to learn throughout their lives. By reinforcing such stereotypes, we discourage students viewed as having deficits from persisting. Identifying strategies that enable a more generative and productive view of disparities in performance across all groups inevitably will advantage everyone.

We have seen recently as political groups promote legislation to end affirmative action that many businesses and industries respond with advertising campaigns to retain these programs because they have found that by enhancing the diversity of their workforce, they also enhance the quality of their product. We are going one step further to suggest that we need to mend affirmative action in important ways, rather than ending it. We need to modify programs so they enable diverse individuals to participate in fields where they have been traditionally underrepresented without continuing to label these individuals as in need of remediation.

Lifelong learning involves a complex set of skills, concepts, capabilities, and content knowledge across varied fields. Inevitably, assessments are much narrower than the accomplishments that they are intended to measure. Unintended consequences of narrowing assessment need diligent attention since they can lead to bias and to serious inequality of opportunity. Decisions about question format, question content, and the alignment of instruction with assessment have far reaching impact for the opportunities of

success among different groups. The complexity of the educational system makes it impossible to anticipate all the consequences of innovation and change in testing and instruction. Rather, what we need are mechanisms for monitoring these impacts and continuously improving all aspects of education.

Test bias is both a product and a concern of the whole system of education in any culture or nation. A test might be biased for one group and not for another. Tests that predict success for one group might not predict success for another. Educational programs may ameliorate weaknesses for one group but not for another. As a result, we need to look at the question of test bias by examining the full educational program.

At the same time, the recognition that test bias is a product of a complex system must not stymie us in our efforts to ensure equity and fairness for individuals. When individuals take advantage of the complexity of the educational system in order to promote the opportunities for some and demote the opportunities for others, serious consequences of a societal nature result. As this paper has illustrated, for example, individuals might use a multitude of indicators to make decisions, yet apply one indicator for one group of individuals and another for a different group, in order to deselect those not desired in the final pool. Furthermore, a single indicator that advantages one group over another—for example, essay questions or multiple choice questions—might be used to support a deficit model or reinforce stereotypes. Designing educational policies to take test bias into consideration and to make effective testing and prediction decisions requires a careful understanding and analysis of potential sources of bias, as well as willingness to

cut through the complexity and look specifically at those factors likely to lead to the most effective solutions to educational problems.

Inevitably, policy makers and observers disagree when tests reveal inequalities. When particular groups fall behind, some argue this is because the educational system is failing them, and others argue that this is a natural and desired outcome for the group. Most commonly, both of these are somewhat true. However, unless we can ensure that particular groups are underrepresented for good reason, it is imperative that we demonstrate evidence that they are not underrepresented because of stereotype or bias.

## FURTHER READING

AAUW Educational Foundation Commission on Technology, Gender, and Teacher Education. (2000). Tech-Savvy: Educating Girls in the New Computer Age. American Association of University Women, Washington, DC.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). Standards for Educational and Psychological Testing. American Educational Research Association, Washington, DC.

Bergmann, Barbara R. (1996). In Defense of Affirmative Action. Basic Books, New York.

Bransford, John D., Brown, Ann L., and Cocking, Rodney R. (Eds.). (1999). How People Learn: Brain, Mind, Experience, and School. National Academy Press, Washington, DC.

Caplan, Paula, Crawford, Mary, Hyde, Janet Shibley, and Richardson, John T. E. (1997). Gender Differences in Human Cognition. Oxford University Press, New York.

Gould, Stephen J. (1981). The Mismeasure of Man. W. W. Norton & Company, New York.

Heubert, Jay P. and Hauser, Robert M. (Eds.). (1999). High Stakes: Testing for Tracking, Promotion, and Graduation. National Academy Press, Washington, DC.

Linn, Marcia C. and Hsi, Sherry. (2000). Computers, Teachers, Peers: Science Learning Partners. Lawrence Erlbaum Associates, Mahwah, NJ.

Linn, Marcia C. and Kessel Cathy. (1996). Success in Mathematics: Increasing Talent and Gender Diversity. In Alan Schoenfeld, Ed Dubinsky, and James Kaput (Eds.), Research in Collegiate Mathematics Education II (pp. 101-144). Providence, RI: American Mathematical Society.

Steele, Claude. (1997). A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance. American Psychologist 52(6), 613–629.

Sternberg, Robert. (2000). The Holey Grail of General Intelligence. Science 289, 399, 401.

U.S. Congress, Office of Technology Assessment. (1992). Testing in American Schools: Asking the Right Questions, OTA-SET-519. U.S. Government Printing Office, Washington, DC.

Valian, Virginia. (1998). Why So Slow?: The Advancement of Women. MIT Press, Cambridge, MA.

Willingham, Warren W. and Cole, Nancy S. in collaboration with Brent Bridgeman. (1997). *Gender and Fair Assessment*. Lawrence Erlbaum Associates, Mahwah, NJ.