Advancing Women in Leadership Online Journal
Volume 21, Fall 2006

What's Really at Stake with High-Stakes Testing?

Dr. Barbara Polnick and Dr. Dianne Reed

Sam Houston State University

High-stakes testing has the potential to make significant differences in the lives of children and adults across a variety of educational settings. Decisions related to high-stakes testing can affect individuals and systems at the local, regional, state, and national levels (Ysseldyke et al., 2004). Advocates of high-stakes testing hope test scores will prompt schools and organizations to reform policy, encourage teachers to adopt more effective practices, and motivate students to work harder based on the school's accountability report (Stecher, 2002). However, some studies indicate that high-stakes testing can have negative as well as positive consequences and that these consequences sometimes result in different experiences for females and minorities. In this article, we define high-stakes tests, describe both positive and negative consequences of these assessments as they relate to females and minorities, discuss reasons for performance differences among diverse groups, and share cautions regarding uses of high-stakes tests.

High-Stakes Tests Defined

"High-stakes" tests are tests that have direct consequences for individuals, groups, or organizations, in which the stakes are high. In public schools, for example, teachers, principals, and superintendents are sometimes warned that their raises, bonuses, or even their jobs may be on the line if students are not successful on these tests. High-stakes tests have the potential to affect opportunities for students by impacting milestones such as graduations, promotions, and admissions into programs of colleges and universities. These tests, once largely used as monitoring devises in educational settings, can also be defined in terms of their direct impact on and changes to a campus, district, or organization as a whole (Carnoy, Elmore, & Siskin, 2003). High-stakes tests include, but are not limited to, state-wide assessments; assessments used to measure "adequate yearly progress" (AYP) as required by No Child Left Behind (NCLB) federal legislation (U.S. Department of Education, 2002); college entrance exams; certification and licensure exams like those used in education, business, medicine, and legal professions; and some aptitude exams. Depending on how they are used, some psychological and intelligence assessments could also be considered high-stakes assessments when used to "sort and select" individuals for specific programs. The use of high-stakes testing is what determines whether the assessment is "high-stakes"-not the design of the test.

Use of High-Stakes Assessments

High-stake assessments have varied in how they have been used over the years. The current wave of assessment-based school accountability combines two traditions in American education: public accountability and student testing. In the past, when accountability and assessment were only loosely connected, assessments were used mainly to divide students into academic tracks, or for diagnostic purposes. Most recently, high-stakes assessments have been used to measure the performance of organizations against a designated set of standards, thus holding individuals, groups, and schools all accountable for student performance (Carnoy, Elmore, & Siskin, 2003). As accountability measures, these assessments have resulted in both positive and negative consequences.

Positive Consequences

New accountability measures based on high-stakes testing are designed to improve the quality of education and provide equity in instruction (Stecher, 2002). By requiring states, for example, to assess and report student outcomes by diverse groups, emphasis is placed on all students achieving standards regardless of their socio-economic status, race, ethnicity, or gender. Changes in school policies designed to make schools more effective in increasing student achievement can increase motivation on the part of the students (Stecher, 2002). For schools, incentives to perform result in support from faculty to be successful, and praise and recognition all serve as motivators for individual students. When performance on high-stakes tests are analyzed for individual students, focused time and effort directed toward diverse population groups passing provides additional support to succeed. It is hoped that with increased pressure to help all groups perform better on the tests, gaps in performance between and among diverse population groups will diminish. High-stakes assessments that promote mastery of national standards may focus curriculum reform efforts toward meaningful content and research-based instructional practices. Meeting those curriculum standards forces students to participate in coursework of the same level(s) in order to pass the exams. For example, all students (girls and boys) taking Algebra II, Geometry, and Calculus increases the likelihood that girls as well as boys will have the necessary pre-requisites for admission to prestigious universities. Females, then, are necessarily encouraged to take the same higher level coursework as males because, unless they do, they will not pass the exams. This is especially important in states where accountability ratings are based on college preparatory exam results such as those offered by the Advanced Placement (AP) testing program.

Negative Consequences

There are several reasons to be concerned about the consequences of high-stakes tests for schools and individual students. In PK-12 schools where high-stakes assessments are associated with accountability systems, use of these assessments can result in rewards and/or sanctions for students and schools. These high-stakes assessments, depending on how they are used as accountability measures, have the potential to negatively impact schools and other organizations.

Increasing the gap . Negative consequences may include an increased gap between low-and high-performing schools, reduced opportunities for females and minorities as well as students with disabilities, and decreased support for national curriculum reform efforts.

Schools whose students fail to pass such tests in sufficient numbers are sometimes sanctioned by the state or national government, which can result in a change in faculty and/or leadership, or a total reorganization of the school. When schools are held equally accountable for meeting performance standards, gaps between schools (the "haves" and the "have nots") may increase as sanctions are imposed because the negative effect for low performing schools and students is greater than the negative effect on high performing schools and students (Shepard, 1991; Carnoy, Elmore, & Siskin, 2003).

Graduation and completion. Where high-stakes assessments translate into high school "exit tests" that all students must pass in order to graduate, completing or not completing school can result in students being marginalized from the mainstream, and becoming a social liability. This action can have significant consequences in high schools where low-income young people-many of them African-Americans and Latinos-"make it or break it" educationally (Mcdermott & Mcdermott, 2002). High-stakes tests impact students with disabilities when high passing standards make graduation as a reward nowhere in sight (Mcdermott & Mcdermott, 2002; Ysseldyke et al., 2004).

Reform efforts . States and national organizations may be affected by international assessments in a way that can significantly impact curriculum reform efforts-efforts which are sometimes put into place to reduce gaps and achieve equity. For example, the National Assessment of Educational Progress (NAEP) is an international exam that has no direct consequences for how well individuals, schools, or even states do on the test. However, this test compares how educational programs in the United States rank when compared to other countries. It also compares how well males and females or different minority groups perform on specific sub-sections of the test. When this information is made public, different factions may use (or misuse) the information to promote or discredit current reform efforts and educational practices supported by various political groups. Practices that have been put into place as a means to achieving equity in education may be discounted by the reporting of a lack of success on this high-stakes assessment.

Benchmarking and narrowing the curriculum. Another negative consequence of high-stakes assessments occurs when schools design instruction to eliminate gaps by focusing instruction on the weak areas of students-not their strengths (Mcdermott & Mcdermott, 2002) . This type of instruction can sometimes negatively impact student performance through a continuous focus on students' failure(s) to perform well, which may serve as a catalyst for giving up or even dropping out of schools and/or universities. Repeated testing, commonly known as "benchmarking," can also negatively impact students' efforts to do their best. This may be especially important for females who research indicates respond differently to pressure in anxiety-related situations (Felson & Trudeau, 1991; Gierl & Rogers, 1996; Pappamihiel, E., 2001). This can affect female students, for example, who may develop anxiety behaviors (which also interfere with their potential to do well) if undue pressure is placed on them to perform ("If only you had gotten three more right on the test, our school would have received a higher rating").

Additionally, achieving equity through increased opportunities may not occur because a minimalized curriculum and/or instruction focused solely on state or local assessments can reduce students' exposure to broad areas of knowledge often assessed on national or international exams, the same exams used for entry into educational programs. In other words, "You can get out, but you can't get in!" The scope of the curriculum is narrowed when there is a focus on national standards to the point that there are few opportunities for creative, innovative thinking, such as those needed for the fields of technology, leadership, space, and medicine. Because creative opportunities can open doors for more students who may not be successful in traditional settings, diverse population groups can be negatively impacted by their elimination.

Student course placement . According to Ysseldyke et al. (2004), using test results to place students in special programs does not always result in increased educational equity and accessibility to content, but can often lead to the development of de facto tracks that students are placed into once they fail high-stakes exams. These "tracks," in which students are assigned, usually on the basis of perceived achievement of skill level, often result in separate schools or programs, classes within grade levels, groups within classes (at the elementary level), and courses within subject areas (at the secondary level) (National Research Council, 1999). For females and some minority students, judgments about perceived achievement based on test results may not be accurate and, therefore, may result in these groups of students being inappropriately placed into courses and/or counseled out of some career opportunities.

Misuse of the mean. A danger in using high-stakes testing occurs when data from these assessments are analyzed and reported in terms of mean scores across population groups and findings are generalized for these groups which can include incorrect assumptions about females, African-Americans, immigrants, etc. (Caplan & Caplan, 1997). For example, even though there are a range of scores across sections of an exam in which some females scored higher than some males, if females are reported at scoring lower (on average) on the quantitative section of the Graduate Records Examination (GRE), stereotypes about females and their abilities to do well in mathematics, engineering, and chemistry may develop. Generalizing that females will not do as well in these areas can and does affect recruitment efforts, program design, as well as faculty and staff selection at universities.

Caplan and Caplan (1997) caution us in *Gender Differences in Human Cognition* that it would be "both nonsensical and immoral to try to restrict flat out the education, treatment and opportunities of individuals simply on account of the phenomenon that the average member of one group scores higher than the average member of the other group" (p. 54). Stereotyping based on scores can also affect perceptions regarding individuals' abilities to "do the job," creating barriers to the selection of certain groups from participating in programs after graduation. Barriers created by the misuse of data reported from high-stakes tests like college entrance exams can, therefore, lead to certain groups being unable to access and succeed in programs that would lead to greater opportunities for success.

Use of assessments to sort and select . In pondering the concept of equity in testing and assessment, it is important to recognize that tests and assessments have traditionally served as sifters which, intentionally or unintentionally, filtered out groups of students from educational opportunities. These groups include, but are not limited to, females, language-minority students, students with low socioeconomic backgrounds, and students of color. This "sort and select" process often depends on the application of the bell-shaped curve, which by design ranks the performance of individuals and sorts them into categories of percentile rankings "below the mean" and "above the mean." Standardized tests like the SAT and GRE tend to especially penalize females and many minority students, where females tend to do worse than males but consistently earn better grades than males. Researchers consistently find that adding test scores to the admissions equations results in fewer females and minorities being accepted than if their academic records alone were considered (Sacks, 2000). For example, historically, the major purpose of college entrance exams has been to predict the success of the first year college performance, but often such scores underpredict what females and minorities can actually do (Aron, Burton, & Poole, 1999). Gender differences in performance on high-stakes assessments used for program admissions can lead, therefore, to an "under prediction" of how well certain groups might do if they were otherwise selected for participation. This underprediction can diminish educational opportunities by denying females and minorities admission to the most competitive colleges and millions of dollars in scholarship money.

Reason for Performance Differences among Diverse Groups

Researchers have found several reasons that account for gender differences in test performance (Langenfeld, 1997). Some studies have found that the context in which test questions are embedded can make a difference in the test performance of females and males (Langenfeld, 1997; Wendler & Carlton, 1987). Females often score higher on questions about relationships, and males score higher on questions about science, sports, and the stock market (Langenfeld, 1997). Educational Testing Systems (ETS) researchers Wendler and Carlton (1987) found that females perform better on test questions that are specific and concrete, as found in questions about science and practical affairs. Designers for ETS had not been successful in completely balancing verbal content with equal references to areas that interest each sex (Wendler & Carlton, 1987).

Another reason why gender differences in test performance might exist relates to the long term effects of stereotyping. Steele and Aronson (1995) found in their research that stereotyping can ultimately affect the performance of individuals for whom society has low expectations. A female who is taking a test in math, but who does not think she is supposed to do well, may become frustrated when faced with a very difficult math problem, thinking "Why try when I can't do this anyway?" and ultimately may withdraw (Steele & Aronson, 1995). According to these researchers, stereotyping can especially impact the test performance of individuals who want to do well in a subject and for whom doing well is very important. Steele (1997) found that in these incidences, stereotyping resulted in stress behaviors which manifested themselves as high blood pressure, anxiety, and/or reading difficulties (re-reading, second guessing, and going back and forth).

Anxiety is another factor which may explain gender performance differences. In a study of 1,112 students in a coaching class taking a practice SAT, it was discovered that two and a half times as many females as males said they were "extremely anxious," and they scored lower on the math SAT than the "somewhat anxious" females (Steele, 2004). Another reason for the gender differences in test performance may be explained by cognitive styles. Females are less likely to be risk takers and to guess at the right answer largely because of their socialization and early education (deNuys & Wolfe, 1985; Sadker & Sadker, 1985).

Time pressure may be an additional reason for gender performance differences. In a recent study conducted by the National Commission on Testing and Public Policy (NCTPP) (1990), students were not given time limits on the SAT. When Kathleen Kelly-Benjamin interviewed a sampling of 20 girls and 20 boys who took the untimed test about their approaches to test taking, she found differences in their strategies. The girls were more likely to work mathematics problems as they had been taught in the classroom; i.e., in a step-by-step manner using formulas or knowledge of geometry; whereas, the boys were more likely to use test-taking strategies such as substituting answer choices in the problem to see which one worked or noticing the answer choice that was different from all the others (usually the correct one) (NCTPP, 1990). Giving the girls all the time they needed seemed to work in their favor but not in the favor of the boys who answered fewer questions correctly despite the expanded time frame.

Cautions in Using High-Stakes Assessments

The American Educational Research Association (AERA) and the American Psychological Association (APA), in conjunction with the National Council on Measurement and Education (1999), developed the *1999 Standards for Educational and Psychological Testing* . These standards represent assessment principles designed to promote fairness in testing and avoid negative impact from the use of assessments. Five standards specifically relate to the development, selection, and use of assessments to minimize potential negative effects on gender, ethnicity, race, socio-economic status, English language learners, and students with disabilities. These include:

• Decisions about students' education, such as retention, tracking, or graduation, should not be based on the results of a single test, but should include other relevant and valid information.

• There should be evidence that the test addresses only the specific or generalized content and skills that students have had an opportunity to learn, when test results substantially contribute to decisions made about students' promotion or graduation. For tests that will determine a student's eligibility for promotion to the next grade or for high school graduation, students should be granted multiple opportunities to demonstrate mastery or materials through equivalent testing procedures.

• School district, state, or other authority mandated tests should clearly describe the ways in which the test results are intended to be used. It is also the responsibility of those who mandate the test to monitor its impact, particularly on racial and ethnic-minority students or students of lower socioeconomic status, and to identify and minimize potential negative consequences of such testing.

• Special accommodations for English language learners may be necessary to obtain valid test scores. If English language learners are to be tested in English, their test scores should be interpreted in light of their limited English skills.

• Special accommodations may be needed to ensure that test scores are valid for students with disabilities. Not enough is currently known about how particular test modifications may affect the test scores of students with disabilities; more research is needed (AERA, APA, NCME, 1999).

Leaders are encouraged to implement the above standards as they incorporate high-stakes testing into their programs. In so doing, they can facilitate equity for all diverse population groups and eliminate some of the misuses of high-stakes testing often found in educational settings today.

Conclusions and Recommendations

In summary, high-stakes assessments have the potential to both positively and negatively affect females and minorities due to several factors, depending on how the results are used. Focusing on performance and expectations for all students can raise the level of awareness about individual students' needs. In such cases, high-stakes assessments can focus instruction and curriculum towards mastery of a unified set of standards. However, while high standards and expectations motivate some students to perform well, others find this type of pressure distracting and debilitating.

High-stakes assessments can be an important tool in evaluating and improving educational programs. However, it is only a part of the formula for quality. When tests are used in high-stakes circumstances, it is important that all involved be aware of and follow the *Standards for Educational and Psychological Testing* in order to ensure that certain groups of the population of test-takers are not disadvantaged by the tests. Such considerations may include gender- and culturally-biased language, test design, and testing parameters. Furthermore, users of high-stakes tests need to recognize that factors such as test anxiety may produce scores that are not representative of an individual's full potential.

High-stakes assessments can negatively impact educational outcomes when stereotyping occurs. This occurs when the results of assessments are generalized and published for the general population to the point that they become a truth. Furthermore, gender differences in performance outcomes on high-stakes assessments used for program admissions can lead to an "underprediciton" of how well females (and minorities) might do if they were otherwise provided the opportunity to participate. In addition, the nature of high-stakes assessments can create anxiety and pressure on people who are already feeling uncertain about their abilities, which can then interfere with how well they perform, producing an inaccurate measure of what they can do. This then can lead to a masking of the true potential of groups of people which are then generalized and purported as truth, perpetuating the stereotyping of individuals and their abilities.

If tests are going to be used to determine which students will advance or be accepted into advanced programs, which courses and programs will be offered, and which type of instruction will be promoted, it is imperative that we understand how to effectively measure student learning and performance. In addition, educators should explore how best to use high-stakes assessments to impact student drop-out/completion rates, graduation rates, course content and design, levels of student anxiety, and instructional practices. Because the stakes can be so high for so many, additional research should begin immediately in order to learn more about the intended and unintended consequences of testing in educational decision making (American Educational Research Association, American Psychological Association, & National Council on Measurement and Education, 1999).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement and Education. (1999). *Standards for educational and psychological testing* . Washington, DC: AERA Publications, Inc.

Aron, R. H., Burton , D. N., & Poole , D. A. (1999). Underprediction of female performance from standardized knowledge tests: A further example from the knowledge of geography test. *Sex Roles: A Journal of Research, 41* (7/8), 529-540.

Caplan, P. J., & Caplan, J. B. (1997). Do sex-related cognitive differences exist, and why do people seek them out? In P. Caplan, M. Crawford, J. S. Hyde, & J. T. E. Richardson (Eds.), *Gender differences in human cognition* (pp. 52-75). New York: Oxford University Press. Retrieved February 19, 2006, from Questia database: http://www.questia.com/PM.qst?a=o&d=57129854

Carnoy, M., Elmore, R., & Siskin, L. S. (Eds.). (2003). *The new accountability: High schools and high-stakes testing* . New York: Routledge Falmer. Retrieved February 19, 2006, from Questia database: http://www.questia.com/PM.qst?a=o&d=108631580

deNuys, M., & Wolfe, L. R. (1985). *Learning her place-Sex bias in the elementary school classroom* . Washington, DC: Project on Equal Education Rights.

Felson, R. B., & Trudeau, L. (1991). Gender differences in mathematics performance. *Social Psychology Quarterly, 54* (2), 113-126.

Gierl, M. J., & Rogers, W. T. (1996). A confirmatory analysis of the Test Anxiety Inventory using Canadian high school students. *Educational and Psychological Measurement, 56* , 315-324.

Langenfeld, T. E. (1997). Test fairness: Internal and external investigations of gender bias in mathematics testing. *Educational Measurement: Issues and Practices, 16* (1), 20.

Mcdermott, D. F., & Mcdermott, T. K. (2002). High-stakes testing for students with special needs. *Phi Delta Kappan, 83* , 7.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America* . Chestnut Hill, MA: National Commission on Testing and Public Policy, Boston College.

National Research Council. (1999). High stakes: Testing for tracking, promotion, and graduation. In J. P. Heubert & R. M. Hauser (Eds.), *Committee on appropriate test use, Board on Testing and Assessment, National Research Council* . Washington, DC: National Academy Press.

Pappamihiel, E. (2001). Moving From the ESL classroom into the mainstream: An investigation of English language anxiety in Mexican girls. *Bilingual Research Journal, 25* , 1 & 2. Retrieved February 10, 2006, from http://brj.asu.edu/v2512/articles/art3.html

Sacks, P. (2000). *Standardized minds: The high price of America's testing culture and what we can do to change it* . Cambridge, MA: Perseus Publishing. Retrieved February 21, 2006, from Questia database: http://www.questia.com/PM.qst?a=o&d=91066127

Sadker, M., & Sadker, D. (1985, March). Sexism in the schoolroom of the '80's. *Psychology Today* , 54-57.

Shepard, L. A. (1991). *Will national tests improve student learning?* (CSE Technical Report 342). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B. M. (2002). Consequences of large-scale, high-stakes testing on school and classroom practice. In L. S. Hamilton, S. P. Klein , B. M., & Stecher (Eds.), Making sense of test-based accountability in education (p. 79). Santa Monica, CA: Rand.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52* , 613-629.

Steele, C. M. (2004). Not just a test. *The Nation* , posted April 15, 2004 (May 3, 2004 issue), 1-4. Retrieved February 10, 2006, from http://www.thenation.com/doc/20040503/steele

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69* , 797-811.

U.S. Department of Education. (2002). *No Child Left Behind Act of 2001* . Retrieved April 6, 2003, from www.ed.gov/offices/OESE/esea/summary.html

Wendler, C. L.W., & Carlton, S. T. (1987). *An examination of SAT verbal items for differential performance by females and males: An exploratory study* . Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Ysseldyke, J., Nelson, J. R., Christenson, S., Johnson, D. R., Dennison, A., Triezenberg,H., et al. (2004). What we know and need to know about the consequences of high-stakes testing for students with disabilities. *Exceptional Children, 71* (1), 75+. Retrieved February 21, 2006, from Questia database: http://www.questia.com/PM.qst?a=o&d=5007601130

Dr. Barbara Polnick is an Assistant Professor in the Department of Educational Leadership and Counseling at Sam Houston State University.

Dr. Dianne Reed is an Assistant Professor in the Department of Educational Leadership and Counseling at Sam Houston State University.