The Relative Equitability of High-Stakes Testing versus Teacher-Assigned Grades: An Analysis of the Massachusetts Comprehensive Assessment System (MCAS)

ROBERT T. BRENNAN
JIMMY KIM
Harvard Graduate School of Education

MELODIE WENZ-GROSS
GARY N. SIPERSTEIN
University of Massachusetts Boston

Which is more equitable, teacher-assigned grades or high-stakes tests? Nationwide, there is a growing trend toward the adoption of standardized tests as a means to determine promotion and graduation. "High-stakes testing" raises several concerns regarding the equity of such policies. In this article, the authors examine the question of whether high-stakes tests will mitigate or exacerbate inequities between racial and ethnic minority students and White students, and between female and male students. Specifically, by comparing student results on the Massachusetts Comprehensive Assessment System (MCAS) with teacher-assigned grades, the authors analyze the relative equitability of the two measures across three subject areas — math, English, and science. The authors demonstrate that the effects of high-stakes testing programs on outcomes, such as retention and graduation, are different from the results of using grades alone, and that some groups of students who are already faring poorly, such as African Americans and Latinos/Latinas, will do even worse if high-stakes testing programs are used as criteria for promotion and graduation.

Harvard Educational Review Vol. 71 No. 2 Summer 2001

Do high-stakes testing programs worsen educational outcomes for racial/ethnic minorities and for girls of all races and ethnicities? In 1994, the editors of the *Harvard Educational Review* argued that, as new methods of evaluating students emerge, a high level of scrutiny should be given to ensure "that new assessment practices do not continue to worsen educational inequality" (Editors, 1994, p. 4). The recent and rapid expansion of high-stakes testing programs has prompted concerns about how these groups of students may be affected when high-stakes tests are used to determine educational outcomes.

In this article, we adopt a definition of "high stakes" established by professional testing and psychological organizations. According to this definition, "when significant educational paths or choices of an individual are directly affected by test performance, such as whether a student is promoted or retained at a grade level, graduated, or admitted or placed into a desired program, the test is said to have high stakes" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 139).

While there are no formal requirements for passing the eighth-grade Massachusetts Comprehensive Assessment System (MCAS) in order to be promoted to the ninth grade, students are nevertheless affected in several ways. First, they are affected indirectly, through sanctions on their schools. The implementation of the MCAS as a system of statewide testing originated in the Educational Reform Act of 1993. The legislation combined mandatory use of the tenth-grade MCAS as a graduation requirement as well as the use of the eighth- and tenth-grade MCAS as an accountability measure for schools; the latter came into use with the first wave of MCAS exams in 1998. School accountability hinges on the percentage of students within a school who are failing or proficient on the exams. Subsequent to each administration of the test, schools were rated and then given targets to meet. Two consequences result from failing to meet the target. First, a school becomes ineligible for a number of funding opportunities, including Title I for exemplary programs. Second, the schools may be granted funds (targeted assistance) to aid them in their efforts to achieve their targets; these funds, however, will only be given to programs approved by the state, including weekend and summer programs for students. Finally, the heavily publicized results are meant to affect choices made by Massachusetts residents, for example, by becoming part of the criteria for choosing a school district to live in, and by brining localized public pressure on underperforming schools to improve (R. Lee, personal communication, June 4, 2001).

Somewhat more directly, principals and teachers are, in some circumstances, under strong pressure to reduce the number of failing scores in their districts. This may result in targeting particular students for remediation or exclusion from more challenging classes. Students and parents alike

may consider a child's score in their decisionmaking about educational issues. In addition, the eighth-grade MCAS can be seen as a "dress rehearsal" for the tenth-grade MCAS, which will be a requirement for graduation, starting with the graduating class of 2003. Students with a poor showing in eighth grade respond in multiple ways, the most desirable being an increased devotion to their studies in order to improve their showing two years down the line, while others, perhaps those already alienated from academics, may actually become discouraged and distance themselves still further from schoolwork. Although there are no formal sanctions for students who do not pass the eighth-grade MCAS, the testing methods are congruent with those used in the tenth-grade MCAS and in many other high-stakes testing programs throughout the country, and as such the eighth-grade results remain quite informative as to how different groups of students fare under such a program.

There are many ways that equitability (Supovitz & Brennan, 1997) might be defined when applied to educational assessments. Because we cannot know the true abilities of each student evaluated, we concern ourselves with the *relative* equitability of high-stakes tests to the prevailing gold-standard measure of school achievement — teacher-assigned grades. According to our definition, if a high-stakes test creates a greater gap between groups as defined by their race/ethnicity or gender, then we would consider the high-stakes test to be less equitable than the prevailing standard. Should high-stakes tests narrow those gaps, we would consider the tests to be more equitable. Because we have no way to assess any student's true abilities, we cannot determine which, if either, of these assessments is more biased. We also cannot choose between competing arguments about the validity of these two forms of assessment, such as the notion that the format of standardized tests may discriminate against certain groups of students, or the idea that teachers "compensate," whether consciously or unconsciously, by inflating the grades of students they consider to be challenged.

Many issues concerning the equitability of using a high-stakes assessment in the educational system may go unresolved for some time. Publishers and supporters of standardized testing programs, as well as critics, have long struggled with the question of bias in tests, and such questions show little sign of being resolved soon. While teacher-assigned grades have been the central criterion used in academic decisionmaking, they have seldom been scrutinized for validity, reliability, and bias. Furthermore, there is at present no one way to definitively assess the bias, if any, inherent in any large-scale assessment method. On the other hand, what can be assessed readily is the *relative* equitability of competing assessment methods, such as grades and standardized tests. A meaningful examination of *relative* equitability can be undertaken if one has access to two or more assessments given to the same students at the same point in time (Supovitz & Brennan, 1997). With such

data, which are rarely available to researchers, analysis can reveal whether any known characteristic of a student has greater or lesser effects for a given assessment system compared with another.

We used eighth-grade scores from the first MCAS, administered in the spring of 1998, to compare the relative equitability of standardized tests and teacher-assigned grades. If, as critics of high-stakes tests argue, the MCAS produces more inequitable outcomes, we would expect the test-score disparities to be larger than the teacher-assigned grade disparities when comparing the performance of White students and minority students, and of boys and girls.

The results of our analyses reveal that the equitability effects of the MCAS vary across the three academic subjects (mathematics, English, science) and according to the student characteristics. For African American and Latino/Latina students, although the evidence for the relative harm to Latino/Latina students is just suggestive, the MCAS results are significantly more inequitable than grades in math, though we did not find significant differences in English or science.[1] For Asian students, on the other hand, we found no significant differences in the equitability of school grades and MCAS scores. Consistent with comparisons between grades and other standardized testing programs, despite earning higher grades, girls score significantly lower on MCAS math and science sections than do boys. In English, however, girls outperform boys in both grades and MCAS by a similar margin.

These results suggest that exclusive reliance on standardized test scores to make high-stakes decisions may worsen educational outcomes for minorities, while for girls the eighth-grade MCAS risks adding to the list of factors discouraging girls from pursuing further studies in science and math.

Literature Review

The Rise of Statewide High-Stakes Testing Programs

Since the passage of Goals 2000 and the reauthorization of Title I in 1994, many states have instituted an ambitious series of reforms to upgrade curriculum, instruction, and assessment. These federal policies have encouraged states to establish more challenging content standards, to align assessments with these standards, and to hold schools responsible for helping all chil-

[1] As we will show later, although the relatively large disparity between grades and MCAS scores for Latino/Latina students was not significant at the common $p < .05$ level, it is fairly large in magnitude. Therefore, even with the relatively small number of Latino/Latina students in the sample, and thus limited statistical power to find small and medium-size effects, we believe that this evidence should be given due consideration. Further, in several statistical models that preceded the full model presented in this article, the relative harm to Latino/Latina students of the MCAS as compared with grades was significant beyond the conventional .05 limit. As we will discuss later, in the case of identifying an effect for African American students participating in a desegregation program (METCO), multicollinearity of predictors in these models tends to undermine precision of the estimates. It does so by inflating standard errors, increasing the p-values. Therefore, we do not present models in which race/ethnicity, school dummies, and METCO participation are all included.

dren master a rigorous body of academic knowledge and skills (Cohen, 1996; O'Day & Smith, 1993; Ravitch, 1995). By creating a system of rewards and sanctions that is tied to student performance on large-scale state assessments, policymakers have relied on high-stakes test results to monitor and evaluate each school's progress in meeting achievement targets. In principle, then, test-based accountability policies create incentives for schools to improve teaching and learning (Elmore, Abelman, & Fuhrman, 1996; "Quality Counts," 1999; Vinovskis, 1996).

The widespread adoption of high-stakes testing programs can be attributed to several factors that facilitate policy implementation (Kingdon, 1995). High-stakes testing programs can be rapidly implemented by legislators and dramatically change curriculum and instruction inside the classroom (Darling-Hammond & Wise, 1985; Haertel, 1989; Koretz, Linn, Dubner, & Shepard, 1991; Koretz & Barron, 1998; Linn, 2000). High-stakes testing programs typically cost less than other educational reforms, at least initially. Politicians have embraced high-stakes tests as a mechanism for leveraging the problem of low student achievement and poor school performance (Clinton & Gore, 1992; Hess, 1999; National Governor's Association, 1986; Orfield & Ashkinaze, 1991; Smith, Heinecke, & Noble, 1999; U.S. Department of Education, 1991; Vinovskis, 1999).

Public support for high-stakes testing programs has risen gradually over time (Phelps, 1998, 1999). In 1958, only half the electorate believed that "all high school students in the United States should be required to pass a standard nationwide exam in order to get a high school diploma" (Hochschild & Scott, 1998, pp. 115–116). In 1976, 65 percent supported the idea of a high school graduation exam; by 1996, 87 percent favored the policy. In 1994, a study (Johnson & Immerwahr, 1994) showed that nearly three-fourths of parents believed that students who could not demonstrate proficiency in reading and writing should be denied a high school diploma. However, in a more recent survey (Public Agenda, 2000), an equally large proportion of parents thought it was "wrong to use the results of just one test to decide whether a student gets promoted or graduates" (p. 2). Chronically poor performance on high-stakes tests may erode public support. In Virginia, failure rates on the high-stakes Standards of Learning (SOL) test remain high, especially among low-income, African American, and Latino/Latina students. As a result, the threat of diploma sanctions and the loss of school accreditation may affect a disproportionate number of minority and poor students and their schools (Benning & Mathews, 1999; Mathews, 1999). A recent poll (Mathews & Benning, 2000) found that 51 percent of all registered voters in the state said that the "test is not working," and 43 percent thought it should be "substantially changed."

The political costs of high-stakes testing, however, must be compared to the educational benefits, especially for minority students and those from families with low socioeconomic status. Indeed, some scholars assert that

high-stakes testing programs are a useful and valid means to promote achievement and to provide educational opportunities and resources for struggling students and schools (Coleman, 1998; College Board, 1999; Hirsch, 1996; Smith & O'Day, 1991). Research by Bishop, Moriarty, and Mane (1997) suggests that requiring all students to reach New York's Regent standards in the five core subjects would increase student achievement, high school graduation, and college attendance, and that these requirements would benefit disadvantaged students the most. In several cities and states, minority students have reached state standards when given time, resources, and assistance (Barth et al., 1998; Grissmer & Flanagan, 1998; Grissmer, Flanagan, Kawata, & Williamson, 2000; Toenjes, Dworkin, Lorence, & Hill, 2000). In Milwaukee, for instance, initially low African American pass rates (7% in 1995) on a high school graduation test were followed by improvements, leading ultimately to a 87 percent pass rate by 1996 (P. Barth, personal communication, March 27, 2001; Haycock, 1996).

Other social scientists challenge the argument that high-stakes tests promote educational excellence and equitability. They argue instead that high-stakes tests harm the short- and long-term performance of minority students. Tests that determine high school graduation pose a major obstacle to African American and Latino/Latina students, who are disproportionately represented among students who fail to meet cutoff scores (Kohn, 2000; Kornhaber, Orfield, & Kurlaender, 2001). Historically underserved minority groups, then, are most likely to become discouraged, drop out of high school, and fail to acquire the skills and credentials needed for success in postsecondary educational institutions and the labor market (Catterall, 1989; Figueroa & Hernandez, 2000; Haney, 2000; Madaus & Clarke, 2001; Reardon, 1996).

When educators in low-performing schools are pressured to raise test scores, they often respond by replacing a substantive academic curriculum with test preparation materials, which, according to some critics, have little educational value beyond raising scores (Gordon & Reese, 1997; Hoffman, Assaf, Pennington, & Paris, in press; Koretz, 1988; Linn, Graue, & Sanders, 1990; McNeil & Valenzuela, 2001; Sacks, 1999). Such practices create new educational inequalities by restricting minority students' access to a comprehensive academic curriculum (McNeil & Valenzuela, 2001; McNeil, 2000). For example, in one high-poverty urban district with a large African American and Latino/Latina population, Koretz et al. (1991) found that teachers were narrowing the curriculum to focus only on materials covered by the test. The observed gains on the high-stakes test failed to generalize to other assessments, which measured the same academic content and skills. Similarly, a recent RAND study has shown that comparatively large gains on the high-stakes Texas Assessment of Academic Skills (TAAS) fail to register on the National Assessment of Educational Progress (NAEP) assessment, suggesting that "schools are devoting a great deal of class time to highly specific

TAAS preparation [and that] schools with relatively large percentages of minority and poor students may be doing this more than other schools" (Klein, Hamilton, McCaffrey, & Stecher, 2000, pp. 13–14).

The link between high-stakes testing and gender equity merits scrutiny as well. Through high school and college, girls' grade point averages are equal to or better than boys', although a gender gap favoring boys on standardized tests of mathematics and science emerges around middle school/junior high school and endures, as evidenced by a gap in scores on the mathematics portion of the SAT (Gallagher, 1998; Kimball, 1989). Adolescent girls who perform poorly on math and science tests may, in the long-run, choose to "dis-identify" with these subjects (Spencer, Steele, & Quinn, 1999), forgoing more advanced technical training later in high school and college (Hewitt & Seymour, 1991; Stage, Kreinberg, Eccles, & Becker, 1985; Wilder & Powell, 1989). A new test that widens the gap between girls and boys in science and math during the critical middle school/junior high school years has the potential, depending on the use and interpretation of the test scores, to exacerbate the already higher rate with which girls stop participating in these subjects.

Study Context

The Massachusetts Comprehensive Assessment System

The 1993 Massachusetts Education Reform Act authorized the state Board of Education to establish academic standards in seven core academic subjects, which are outlined in the state curriculum frameworks. Starting in the spring of 1998, a battery of criterion-referenced tests, the Massachusetts Comprehensive Assessment System, or MCAS, was administered to assess student mastery of these standards in grades four, eight, and ten. According to state officials, the goal of these new standards and assessments is straightforward: "The Massachusetts Curriculum Frameworks and MCAS together create a new state system designed to support students, parents, teachers, and schools by uniformly promoting high academic standards for all students of the Commonwealth" (Massachusetts Department of Education, 1998, p. 5). Beginning in 2001, high school students must earn a "passing" score on the tenth-grade English and mathematics MCAS as a requirement for receiving a high school diploma (Goertz, Duffy, & Carlson-LeFloch, 2000).

The most widely used test item format in the MCAS is the multiple-choice question. The math tests also include short-answer questions, which require students to work out a short series of computations and then provide a brief explanation of the steps leading up to the answer. Open-response questions are used in all subjects; such questions ask students to generate a short response in the "form of a narrative or a chart, table, diagram, illustration, or graph" (Goertz et al., 2000, p. 3). On the English examination, students must supply a writing sample. For each test, scaled scores range from 200 to

TABLE 1
Performance Standards on the Massachusetts Comprehensive Assessment System

Level	*Score Range*	*Definition*
Advanced	260–280	Students at this level demonstrate a comprehensive and in-depth understanding of rigorous subject matter, and provide sophisticated solutions to complex problems.
Proficient	240–259	Students at this level demonstrate a solid understanding of challenging subject matter and solve a wide variety of problems.
Needs Improvement	220–239	Students at this level demonstrate a partial understanding of subject matter and solve some simple problems.
Failing	200–219	Students at this level demonstrate a minimal understanding of subject matter and do not solve even simple problems.

Source: Goertz et al., 2000, p. 8.

Note: Special education students are defined as students receiving services under Individual Education Plans.

280 and are subsequently assigned one of four proficiency levels, as shown in Table 1.

In the first administration of the MCAS in 1998, a majority of test-takers scored in the "failing" and "needs improvement" categories. Failing grades on the MCAS were distributed quite unevenly across racial groups. Statewide, on the eighth-grade English MCAS, 8 percent of Whites failed, compared with 25 percent of African Americans, 38 percent of Latinos/Latinas, and 15 percent of Asians. About one-third (34%) of Whites failed the math portion of the exam, while 75 percent of African Americans, 79 percent of Latinos/Latinas, and 36 percent of Asians did not pass. Finally, 31 percent of Whites, 74 percent of African Americans, 80 percent of Latinos/Latinas, and 43 percent of Asians did not pass the science portion (Massachusetts Department of Education, 1998).

High failure rates, among other factors, have provoked strong opposition to the tests. For example, the Massachusetts Association of School Committees, an organization representing local school boards, recently approved a resolution urging state legislators to delay use of the MCAS for graduation (Klein, 2000).

Testing Equitability and Suburban Schools

One problem in trying to make sense of raw data on high-stakes testing is that socioeconomic status and other important characteristics of students are unevenly distributed across schools. For example, the widely touted differences between racial groups on these tests may be attributable partly to

the fact that the children attend different schools. Schools located in inner cities tend to be made up largely of minority students who are of lower socioeconomic status. Even within inner cities there is a confounding of race, socioeconomic status, and other factors. For example, White students with greater economic means may be more likely than poorer White students and minority students to flee the public system and seek an education in private and parochial schools. Within the public system, students may also be unequally distributed due to a system of neighborhood schools, magnet schools, or exam schools.

While no one group of schools may be ideal for disentangling differences between racial/ethnic groups or socioeconomic strata, study of suburban schools offers some advantages. While suburbs continue to be predominantly White, they are becoming more integrated as some minority families gain financial resources and choose to move to the suburbs. Studying only suburban schools helps to offset some of the differences in socioeconomic status that plague studies that attempt to include both urban and suburban schools. In particular, relying on a sample of suburban schools drastically reduces differences in curriculum, resources, teacher qualifications, parental involvement, and other critical features of schools that vary widely between suburban and inner-city schools. By using an exclusively suburban mix of schools, we greatly reduce the threat that effects in our sample are artifacts of students having drastically different school experiences. In comparison to their counterparts in urban schools, students in suburban schools usually experience a more rigorous curriculum, face tougher academic preparation and competition, and enjoy daily interactions with stronger students and more highly skilled and experienced teachers (Eaton, 2001; Kahlenberg, 2000; Orfield, 1996; Rothstein, 1998; Wells & Crain, 1994). Suburban schools tend to be more reflective of the overall composition of the suburban town than neighborhood schools in cities, or magnet and exam schools.

While the restriction of the sample to suburban schools goes a long way toward disentangling the confounding of a student's race/ethnicity and socioecomonics with the strengths and weaknesses of urban and suburban schools, the choice to limit our sample to such suburban schools consequently greatly limits the ability to generalize our findings back to anything other than suburban schools. We believe that the sort of widespread disparities on an entire host of variables that distinguish urban from suburban schools would, at least at this time, prevent us from effectively disentangling the effects of an individual student's race/ethnicity and characteristics of the school attended. Unfortunately, limiting the sample of school districts does not fully uncouple socioeconomics and achievement at the level of the individual student. Further, inequalities between racial/ethnic groups on a national scale make the isolation of socioeconomic factors from race/ethnicity difficult across a broad range of studies in education, public health, sociology, and other disciplines (Massey & Denton, 1993; Orfield, 1996). In some

areas, programs attempt to bring minority students from communities in the lower socioeconomic strata to school districts in the higher strata.

Unfortunately, although we have data for each student in the sample, in particular the combination of both teacher-assigned grades and MCAS scores, which together form the requisites for an analysis of relative equitability, we do not have data at the level of the individual student on socioeconomics, or a reasonable proxy, such as free and reduced-price lunch status. The consequence of this is that we cannot determine whether any racial/ethnic differences in relative equitability should be attributed to a student's race/ethnicity or to differences in personal resources, although we are able to explore this relationship somewhat by comparing in a limited way bused-in urban students to minority students living within a suburban district. We do not believe that this distinction is important, however, because we do not argue that observed differences between racial/ethnic groups should be attributed to the student's race/ethnicity per se. Since economic differences between racial/ethnic minorities and the White majority persist across many types of communities (as we have argued above), including suburbs, we believe that examining the question of relative equitability of teacher-assigned grades versus high-stakes test scores in typical suburban communities in which racial/ethnic groups may be stratified by differences in economic resources will yield results that are broadly generalizable to most suburban communities.

Suburban schools are expected to perform better than schools in the central city and to improve educational opportunities for minority students (Chubb & Moe, 1990; Jackson, 1985; Miller, 1999). An econometric analysis by Kain and O'Brien (2000) suggests that between 12 and 30 percent of the African American–White test-score gap on the high-stakes TAAS could be eliminated by sending African American students living in the inner city to suburban schools. Suburbs, however, are not a panacea for inequality. As suburban communities become increasingly heterogeneous (Richard, 2000), many districts are struggling to narrow minority achievement gaps on both standardized tests and teacher-assigned grades (Fletcher, 1998; Levine & Eubanks, 1990; Ogbu, 1994).

Methods

We are able to undertake an analysis of relative equitability because we have data on both teacher-assigned grades and MCAS scores in English, math, and science from a sample of eighth-grade students. We also have identifiers for race/ethnicity and gender. Our data come from a sample of suburban schools, which offers us the opportunity to investigate the magnitude of the differences among the various groups of students when there are not staggering differences in the socioeconomic resources of the students and the schools they attend. Further, we are able to employ a statistical technique to

remove school-level differences in MCAS achievement and GPAs across the six schools in our sample.

We use scores from the first year (1998) of MCAS. First-year scores are unique because practices that aim solely to boost scores — test coaching, teaching actual test items, narrowing curriculum and instruction, and other instructional practices — are minimal (Linn, 2000). Because "test-wiseness" is often unrelated to a student's knowledge of academic skills and knowledge, its influence should be reduced in order to make valid inferences of student achievement possible (Millman, Bishop, & Ebel, 1965; Prell & Prell, 1986).

Sample and Measures

We chose to study eighth-grade performance because it sums up years of learning and predicts high school achievement trends. Studies using the National Educational Longitudinal Survey (NELS) data support the importance of studying students in middle school. Analysis of NELS by Phillips, Crouse, and Ralph (1998) suggests that test-score "divergence between Blacks and Whites with initially similar skills seems to occur before high school" (p. 257). Another study of NELS reaches a similar conclusion: White students score higher on eighth-grade math and reading tests than African American and Latino/Latina students, and these test-score gaps remain similar throughout high school (National Center for Education Statistics, 1997). While the performance of boys and girls on standardized tests remains virtually the same in elementary school, boys begin to do better, particularly in math and science, in middle and high school (Hyde, Fennema, & Lamon, 1990; Kimball, 1989; Stage et al., 1985). Therefore, eighth-grade performance has important implications for gender equity because the gaps in math and science achievement begin to emerge at this critical stage of adolescence.

Students in this analysis are part of a longitudinal study examining youths' adjustment in middle school. Included in this sample are 736 eighth-grade students (371 boys and 365 girls) drawn from six public middle schools within four suburban school districts near Boston.[2] All students in the se-

[2] Our sample originally included fifty-four students with learning problems who were receiving special education services; nine of these students were in self-contained classrooms, while the remainder attended at least some classes with general education students. Of the fifty-four students, we had MCAS scores for forty-eight. We do not have data to determine whether all eight students who did not take the MCAS were from self-contained classrooms; further, and more critically, we do not have information as to which, if any, of the forty-eight students taking the MCAS were offered and made use of accommodations, or what those accommodations might have been. Recent work such as Fuchs et al. (2000) and Kleinert, Haigh, Kearns, and Kennedy (2000) has had more complete data on special education students and vastly larger samples than were available to us. We did undertake an analysis including the forty-eight students with MCAS scores, finding that, in all three subjects, the MCAS scores were substantially and significantly less equitable than grades in the same subject. Our primary recommendation in light of such a finding would have been that system-wide procedures for determining and implementing valid and appropriate accommodations be put in place. Most thinking on this issue since the first administration of the MCAS in 1998 has reached a similar conclusion (see Kleinert et al., 2000, for more discussion).

lected schools were included in the sampling frame, and parental consent was solicited. All students whose families consented were included in the data collection and in the analysis. Starting in the spring of fifth grade and continuing through the spring of eighth grade, we surveyed students and gathered data from school records in the fall and spring of each school year. Each student's eighth-grade spring grades in English, math, and science were obtained from report cards. Test scores came from results from the MCAS battery taken that spring. Race and gender information was gleaned from school records. As a requisite step toward an analysis comparing the relative equitability of teacher-assigned grades and MCAS scores, both assessments were transformed to be on the same scale, which had a mean of 0 and a "true score" variance of 100, which is the equivalent of a standard deviation of 10 (see Supovitz & Brennan, 1997, for details).

Within the sample, 4 percent of the students are African American (n=36), 5 percent are Asian (n=47), 2 percent are Latino/Latina (n=15), 87 percent are White (n=784), and 3 percent are from other ethnic backgrounds (n=25), including mixed race/ethnicity and racial/ethnic groups other than White, African American, or Latino/Latina.[3] The minority percentages in this sample closely reflect the percentages of minority students attending the schools from which the sample was drawn. In two of the districts the minority percentage in the schools is somewhat greater than the percentage within the community because the districts participate in the Metropolitan Council for Educational Opportunities (METCO) program, which brings students from the inner city to suburban schools.[4]

Table 2 shows district-level demographic information for the four school districts involved in this study. The communities represented in Districts 1, 2, and 3 in Table 2 are labeled by the U.S. Census as economically developed suburbs. Although individual family income data are not available for students in the sample, the median household incomes of the communities in these districts are available and range from $53,488 to $59,719. Each of these districts has one middle school fed by three to five elementary schools. Each middle school in these districts houses approximately six hundred to seven

[3] Unfortunately, further information about the race/ethnicity of students classified as "other" by their school districts was not available. This group made up 3 percent (n = 25) of the sample. We did not want to exclude this group just because we were not certain as to their race/ethnicity. In the wake of the 2000 U.S. Census, more fine-grained categories of race and ethnicity are emerging and will, no doubt, be available to future researchers of school equity. Our analysis shows that, regardless of the precise definitions of this segment of the sample, students classified as "other" appear to share some of the disadvantages seen in the African American and Latino/Latina groups. Considering this group demonstrates that inequities in educational outcomes are not limited to the groups we have long focused on, African American and Latino/Latina, but are shared by other groups, such as bi- and multiracial children, that are a growing segment of the population of America's children, and which certainly are among the children classified as "other" by the school districts in this study. Nonetheless, since we do not know the composition of this group we do not discuss the results in relation to literature on disadvantage.

[4] In the two communities participating in the METCO program, 100 percent of the African American students were participants. No other students in these two districts were METCO participants.

TABLE 2
Median Income, Per Pupil Expenditures, Percentages Receiving Free or Reduced Lunch, and Percentages of Students' Race/Ethnicity by School District

	School District			
	1	2	3	4
Median Household Income	$53,488	$53,492	$59,719	$38,859
Per Pupil Expenditures ($) (SpecED/GeneralED)*	10,364 / 5,499	7,911 / 5,455	13,530 / 6,464	13,862 / 5,313
% Receiving Free/Reduced Lunch	5.2%	3.8%	5.4%	16.7%
Racial/Ethnic Composition				
% African American	3.5%	4.3%	5.6%	10.9%
% Asian	7.7%	3.5%	9.1%	3.6%
% Latino/Latina	1.6%	1.7%	2.2%	2.9%
% Native American	0.1%	0.2%	0.1%	0.1%
% White	87.1%	90.4%	83.0%	82.4%

* District expenditures are broken down by per capita costs for special education and general education students separately. In general, this provides for more meaningul comparisons across districts because expenditure rates across all students will be inflated by the proportion of students within a district who receive special services.

hundred sixth, seventh, and eighth graders. All operate on a middle school model, with students in each grade broken down into smaller teams. District 4 is labeled by the U.S. Census as an urbanized center. The median household income for this community is $38,859. This district has three middle schools, two of which are smaller schools connected to or next to one of its primary feeder elementary schools. Thus, two of the three middle schools in District 4 are more like K–8 schools. Each of these two smaller middle schools houses approximately three hundred sixth-, seventh-, and eighth-grade students coming from three different feeder elementary schools, respectively, while the larger middle school has approximately 550 students coming from five feeder elementary schools. This larger middle school also runs on a middle school model with smaller teams within each grade. District 4, which does not participate in the METCO program, has a slightly higher percentage of African American and Latino/Latina students than the other three districts (Table 2) and a lower socioeconomic status across the community. However, while District 4 has a slightly higher percentage of Latino/Latina students and a higher percentage of African American students than the other schools in the sample, more than 80 percent of students in District 4 are White. In the "Results" section, below, the middle schools in Districts 1, 2, and 3 are referred to as Schools 1, 2 and 3, respectively, while the three schools in District 4 are referred to as Schools 4, 5, and 6.

Analytic Overview

Determining the relative role of predictor variables — specifically, race/ethnicity and gender on standardized tests scores on the one hand, and teacher-assigned grades on the other hand — requires modeling these outcomes simultaneously (Supovitz & Brennan, 1997). An approach using separate ordinary least squares regression models will not be able to test differences in the effect of predictors, because tests for differences between separately estimated regression coefficients depend on the assumption of independence of samples (Cohen & Cohen, 1983). In the instance where a standardized test score and a teacher-assigned grade are given for the same student, the scores are assumed to be correlated. The present data provide an additional and interesting challenge. Not only are there MCAS scores and teacher-assigned grades for each student, but there are scores and grades for each of the three subjects as well. While each pairing of MCAS scores and grades within a subject area (e.g., history) could be modeled separately, modeling all three disciplines in a single model offers the advantage of being able to estimate the correlations within a discipline and across assessment methods as well as the correlations across scores within an assessment method, potentially yielding evidence of the validity of the measures.

Two established and expanding data analytic strategies offer the potential to model simultaneously six correlated outcomes per subject. Structural equation modeling (SEM) using latent variables (Asher, 1983; Bollen, 1989; Jöreskog & Sörbom, 1989; Long, 1983) is one approach. Structural equations are extremely useful in modeling correlated outcomes, and they are flexible in modeling the effect of predictors on those outcomes. The analyst also has great flexibility in specifying whether various terms in the model should be treated as correlated or uncorrelated. Another approach is the use of hierarchical linear modeling (HLM) (Bryk & Raudenbush, 1992), also known as multilevel modeling (Goldstein, 1987, 1995). Hierarchical models allow the modeling of multiple correlated outcomes (Barnett, Marshall, Raudenbush, & Brennan, 1993; Raudenbush, Bryk, Cheong, & Congdon, 2000; Raudenbush, Brennan, & Barnett, 1995; Supovitz & Brennan, 1997) in a fashion similar to that of SEM. While SEM allows more control over the assumptions about covariance, HLM is more flexible with regard to the data. Structural equation modeling approaches do not deal directly with unbalanced data; for instance, if a student has only two of the three pairs of scores, or only one element of a pair, HLM allows the inclusion of such a case where SEM does not. Assuming that the underlying theoretical assumption that all the outcomes for a given individual will be correlated is reasonable, HLM will be preferable to SEM when the data are unbalanced or missing. Should the data be clustered within a number of schools, HLM permits inclusion of an additional level of data to account for the correlation of scores within schools that may result from unobserved SES differences, dif-

ferences in school resources, and other school characteristics. While there have been promising developments in SEM to account for correlations within social units (Muthén, 2001), these have yet to be widely employed and documented.

In our analysis, we chose to partially replicate and to expand a method proposed by Supovitz and Brennan (1997) in the *Harvard Educational Review*, which investigated the relative influence of gender, race, and socioeconomic status on standardized reading-test scores and portfolio assessments of language arts. That analysis employed a multivariate outcome application of HLM, in which the results of the two assessments were modeled by simultaneous equations containing the student demographic predictors. Our analytic strategy expands this model by modeling six, rather than two, outcomes per student.

In our two-level HLM, the first level represents the "measurement model" (Barnett et al., 1993; Raudenbush, Brennan, & Barnett, 1995; Supovitz & Brennan, 1997). At this level, each of the outcomes is treated as latent (in a similar fashion to SEM) and free of measurement error, thus the tendency of the relationship between a predictor variable and the outcome to be underestimated due to measurement error in the outcome is checked. This was accomplished by using known reliabilities for each form of assessment. For the MCAS, reliabilities were obtained from the Massachusetts Department of Education (1999), while for teacher-assigned grades we used the correlation between fall and spring grades within a subject area as estimated from the sample. We refer to these estimates as "true score" estimates; that is, estimates free of measurement error, of teacher-assigned grades, and of MCAS scores. Several other methods are available for modeling latent outcomes in HLM, for example, using parallel scores (Barnett et al., 1993; Raudenbush, Brennan, & Barnett, 1995) or modeling each of the original items and weighting them in accordance with item-response theory (Janssen, Tuerlinckx, Meulders, & De Boeck, 2000). These other approaches require access to more than one assessment of the same construct or the original individual items, neither of which was available in this analysis.

The second-level of the model represents the individual student. Included at this level are the race/ethnicity and gender of each student. While the inclusion of five schools in this analysis creates the possibility of a third level to the hierarchical model, the number of the schools is smaller than that usually accepted for an upper level in the model. Instead of specifying a third level, we address the clustering of students within schools using a "fixed effects" approach (Hanushek & Jackson, 1977) by including a series of dummy variables to represent the school attended. This approach removes all differences in school mean values on the six assessments, essentially controlling for differences, whatever the source may be, between schools. By isolating any between-school differences in the mean of teacher-assigned grades and

MCAS scores, we eliminate the threat that variations across students by race/ethnicity are a result of the fact that student race/ethnicity is not evenly distributed across the schools in the sample.

In the present model, each of the six possible outcomes is represented by a 0/1 dummy variable to indicate which assessment is represented. Because the number of dummy variables is equal to the number of observations per student (i.e., six when the data are complete), the model we specify at level one has no intercept term. Thus, the level-one model is as follows:

$$Y_{ij} = \beta_{1j}X_1 + \beta_{2j}X_2 + \beta_{3j}X_3 + \beta_{4j}X_4 + \beta_{5j}X_5 + \beta_{6j}X_6 + r_{ij},$$

where Y_{ij} is the observed value for assessment i (i = 1, 2, 3, 4, 5, 6) for student j. β_{1j} to β_{6j} represent the true-score estimates for student j on each of the six assessments, where β_{1j} is math grade, β_{2j} is English grade, β_{3j} is science grade, β_{4j} is math MCAS, β_{5j} is English MCAS, and β_{6j} is science MCAS, and r_{ij} represents the measurement error. While the measurement error is represented in the equation by the term r_{ij} with variance σ^2, these parameters are, in fact, not estimated by HLM, but, rather, fixed. The reliability, or both the teacher-assigned grades and the MCAS, is incorporated into the model by the weighting of each of the observations according to the reliability of that measure. In contrast, models that incorporate multiple observations for each measure either through the use of parallel subscales (Barnett et al., 1993; Raudenbush, Brennan, & Barnett, 1995) or through the use of individual items with an underlying item-response model (Janssen et al., 2000) estimate σ^2, and the reliability of the outcome is estimated by the model, rather than through the use of an external reliability estimate (see also Supovitz & Brennan, 1997). Variables X_1 to X_6 are 0/1 dummy variables indicating for which assessment the score (Y_{ij}) is observed. As mentioned above, this equation does not contain an intercept (β_0) term. While not directly calculated by the model because we utilize information about the reliability of measures from other sources (published reliabilities of the MCAS and reliabilities of teacher-assigned grades using fall and spring grades), the level-one variance (known as σ^2) represents the variation attributable to measurement error.

Each of the β_{qj} terms, the estimated true score for student j on assessment q, becomes an outcome in a level-two equation. The six (one for each assessment) level-two equations take the following form:

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1} W_{1j} + \gamma_{q2} W_{2j} + \gamma_{q3} W_{3j} + \gamma_{q4} W_{4j} + \gamma_{q5} W_{5j} + \gamma_{q6} W_{6j} + \gamma_{q7} W_{7j} + \gamma_{q8} W_{8j} + \gamma_{q9} W_{9j} + \gamma_{q10} W_{10j} + u_{qj},$$

where β_{qj} is the "true score" estimate on assessment q (q = 1, 2, 3, 4, 5, 6) for student j, W_{1j} to W_{10j} are dummy variables describing student j, γ_{q1} to γ_{q5} are the estimated coefficients for each of the student characteristic dummy variables, γ_{q6} to γ_{q10} are the estimated coefficients for the school dummy variables, and u_{qj} is the error for student j on assessment q. Specifically, W_{1j} is a dummy for African American race/ethnicity (1 = African American, 0 = not

African American), W_{2j} is Latino/Latina race/ethnicity (1 = Latino/Latina, 0 = not Latino/Latina), W_{3j} is Asian race/ethnicity (1 = Asian, 0 = not Asian), W_{4j} is "other" race/ethnicity (1 = "other," 0 = not "other"), W_{5j} is gender (1 = female, 0 = male), and W_{6j} to W_{10j} represent attendance at Schools 1, 2, 4, 5, and 6. Note that White race/ethnicity is the reference category (in other words, when the value of the four dummies for race/ethnicity are all equal to 0). Male is the reference gender (represented when the dummy for female is equal to zero), and District/School 3 is the reference school (represented when the school dummies are all equal to 0). The variance for each of the β terms (τ_{qq}) represents the "true score" variation in scores on each of the three sections of the MCAS and in each of the three teacher-assigned grades for a subject area.

Because we have correctly specified a model for intercorrelated scores at the individual level, we may then use hypothesis tests (Raudenbush, Bryk et al., 2000; Bryk & Raudenbush, 1992) to test differences between the estimated effects of student characteristics on teacher-assigned grades and the MCAS. The results of these hypothesis tests answer the question of the relative equitability of the two assessments.

Another useful set of estimates obtained from the hierarchical models is a correlation matrix of grades and MCAS scores in all three subject areas. Of note is the fact that because the level-one model accounts for measurement error, these correlations are interpreted as estimates of the population "true score" correlations (i.e., the correlations that would be observed if all the instruments were free of measurement error). The HLM estimates of these correlations tend to be more accurate than those obtained through manual correction for attenuation (Carmines & Zeller, 1979; see Raudenbush, Brennan, & Barnett, 1995, for a discussion of the "true score" correlations), which can, in some instances, estimate correlations greater than one. Large correlations between the two assessments (MCAS scores and grades) within a subject would tend to support the notion that the assessments validly assess students' mastery of the content within a discipline. Large observed correlations among the teacher-assigned grades across the subjects and among the MCAS subscores across the subjects are harder to interpret. We would expect fairly large correlations, because students who do well in one subject also tend to do well in other subjects (Jencks & Phillips, 1999); however, large correlations within an assessment method as compared with the correlations between the two assessments within a subject area might also hint that the assessment method is more strongly predictive of a student's performance than mastery of the particular subjects. For example, strong correlations among the MCAS scores could suggest that students who do well because of a quality such as "test-wiseness" do well on all parts of the test, regardless of the extent of their abilities in the subject. Likewise, students who get good grades due to an ability to please teachers may tend to receive high grades across the board independent of their achievement.

TABLE 3
Mean MCAS Scores by Race/Ethnicity and Gender

	English (s.d.)	*Math (s.d.)*	*Science (s.d.)*
Race			
African American	242.50 (7.94)	224.58 (18.36)	225.92 (13.06)
Asian	246.44 (9.49)	246.06 (17.31)	237.89 (16.04)
Latino/a	237.75 (14.36)	228.50 (21.80)	224.75 (19.24)
Other	239.6 (15.24)	227.87 (23.72)	222.4 (16.15)
White	245.48 (11.07)	240.57 (20.35)	236.01 (16.10)
Gender			
Girls (n=329)	246.75 (10.9)	237.85 (21.66)	233.59 (16.68)
Boys (n=332)	243.49 (11.28)	241.73 (19.41)	236.87 (15.76)
All	245.00 (11.26)	239.67 (20.68)	235.14 (16.34)

Results

In this section, we first present descriptive statistics on student MCAS scores broken down by student race/ethnicity and gender. Second, we evaluate the correlations among the sections of the MCAS and among teacher-assigned grades, as well as the correlations between MCAS scores and grades in the same subjects. Third, we introduce the results of our hierarchical models comparing the teacher-assigned grades for racial/ethnic minority students to grades for White students, and grades for girls to grades for boys. Fourth, we furnish results from models that repeat these same comparisons using MCAS scores. Fifth, we present group statistical tests to determine whether the differences among the groups are statistically different across the two assessment methods; in other words, we test the question of whether the gap in equitability is either larger or smaller for the MCAS compared with teacher-assigned grades. Finally, we compare our findings to quantitative findings gleaned from results of the NAEP tests administered in Massachusetts.

Descriptive (Bivariate) Results

One way to investigate group differences is to look at the aggregated scores across all the schools in our sample. While these comparisons reveal important group differences, they introduce a possible confounding of group membership with school attended. While this aggregation may matter little in the case of gender, it may introduce bias in the case of race/ethnicity, which is not evenly distributed across schools. Nonetheless, reviewing the aggregated statistics provides insight into how various groups of students fare on the MCAS when grouping by school is ignored.

FIGURE 1
English MCAS Scores by Race/Ethnicity

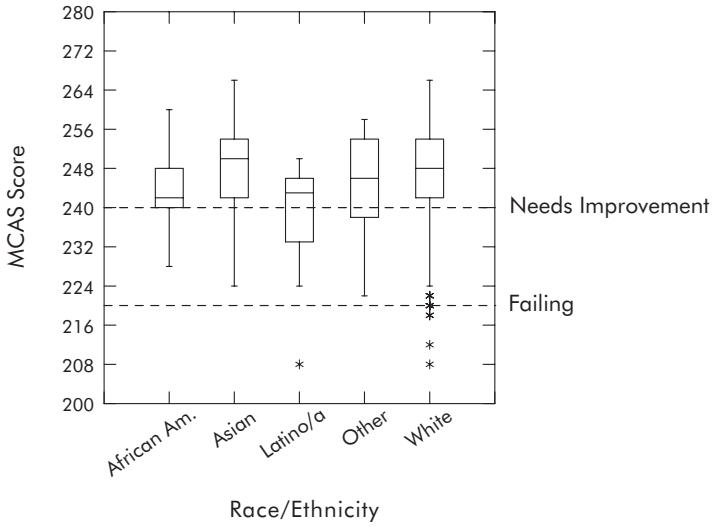

FIGURE 2
Math MCAS Scores by Race/Ethnicity

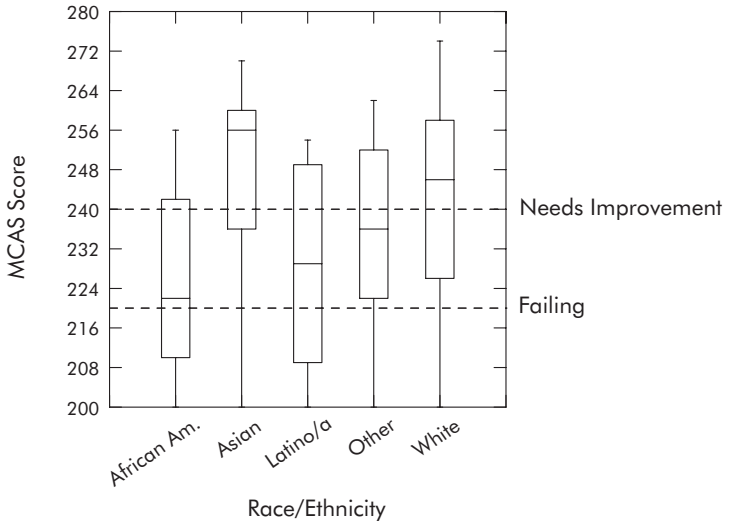

Table 3 displays mean MCAS scores in all three subjects, broken down by race/ethnicity and gender. The distributions for each of the MCAS scores by race/ethnicity and then gender are represented as box plots in Figures 1–6.

Figure 1 is a set of box plots that display the English MCAS scores for students in our sample, broken down by race/ethnicity. These are standard box plots in which the line dividing the box represents the median value for each group; the lower and upper horizontal ends of the boxes, known as "hinges," represent the border of the first and third quartiles, respectively (in other words, the middle 50% of scores lie within the box); the lines or "whiskers" beyond the boxes represent values lying up to 1.5 times the interquartile range beyond the box; and the asterisks represent the values lying more than 1.5 times the interquartile range beyond the box. In these plots, the lower dotted line represents the cutoff score for "failing," while the dotted line above it represents the cutoff score for "needs improvement," as defined by the Massachusetts Department of Education (1998; see Table 1). The mean scores (Table 3) indicate that in the entire sample Whites and Asians earn higher scores in all three subjects than African Americans, Latinos/Latinas, and "others." Figure 1 reveals that the median English score for each of the racial/ethnic groups falls above the level of "needs improvement" (i.e., 240), and that for both Latinos/Latinas and "others" substantial numbers of students fall below the cutoff, whereas few White and Asian students fall below it.

For all of the racial/ethnic groups other than Asian, mean scores on the math portion of the MCAS are lower than on the English portion (Table 3). Figure 2 confirms that many more students fall below the "needs improvement" threshold in math than in English. In general, the scores on the science portion are lower still than on the math portion, with Asians scoring lower in science than in either English or math. African Americans are alone in scoring about the same on the math and science portions (Table 3). African Americans, Latinos/Latinas, and "others" are hard hit in science, with the median science scores lying in the "needs improvement" category and substantial numbers failing (Figure 3).

As expected, based on years of study of middle school/junior high school girls' and boys' scores on standardized tests, girls score somewhat better in English than boys while scoring somewhat lower in math and science. The largest gap in mean scores between boys and girls is in math, where girls average nearly four points less (Table 3). Figure 4 reveals a moderate difference in the proportion of girls and boys falling into the "failing" and "needs improvement" classification (14% of girls, 21% of boys) on the English portion of the exam. In both math and science categories, when compared to the proportions who fall in these categories in English, more students fall in the lowest two categories. On the math portion of the MCAS, substantially fewer boys are categorized into "failing" or "needs improvement"(38% of boys, 45% of girls), as seen in Figure 5. In science, the median score for girls falls

FIGURE 3
Science MCAS Scores by Race/Ethnicity

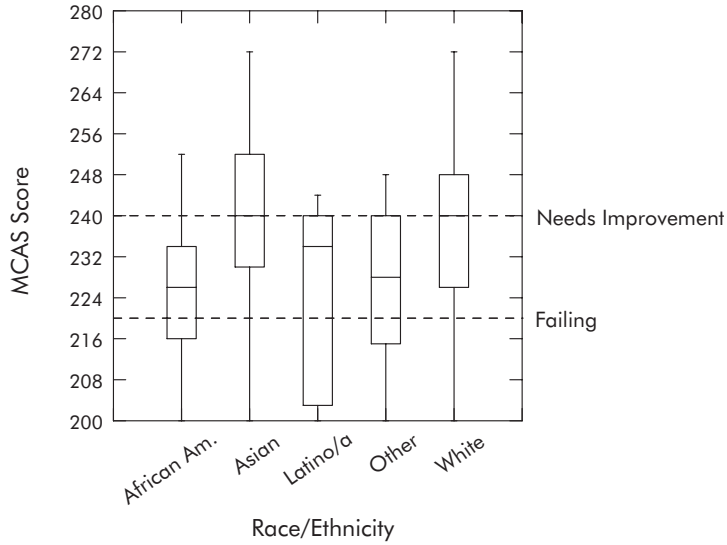

FIGURE 4
English MCAS Scores by Gender

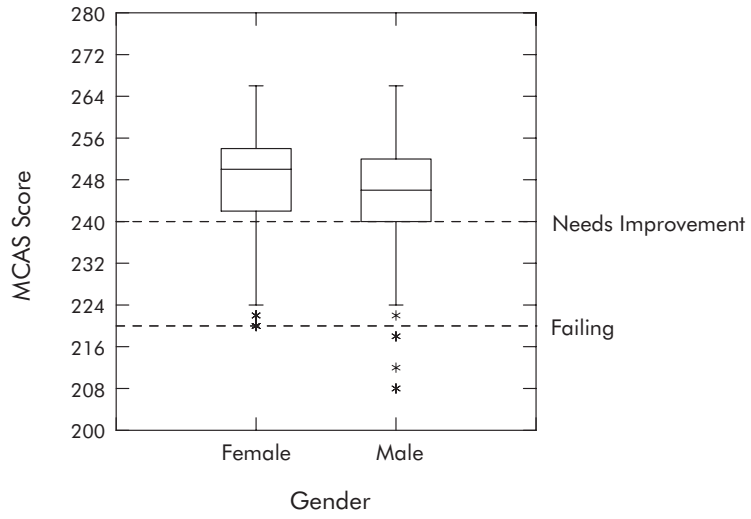

FIGURE 5
Math MCAS Scores by Gender

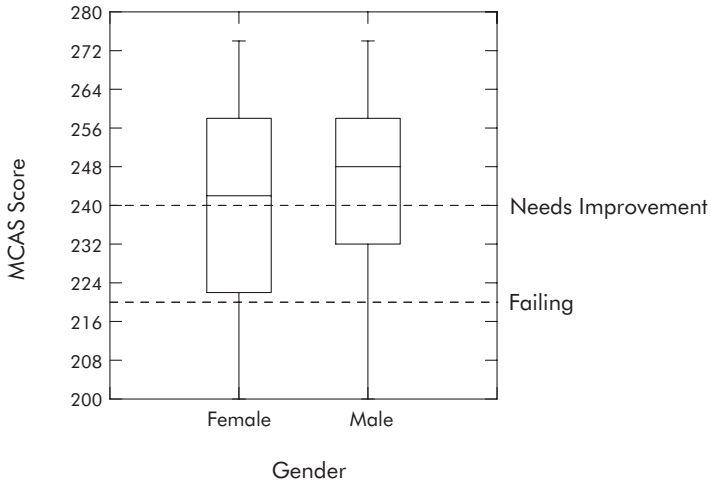

FIGURE 6
Science MCAS Scores by Gender

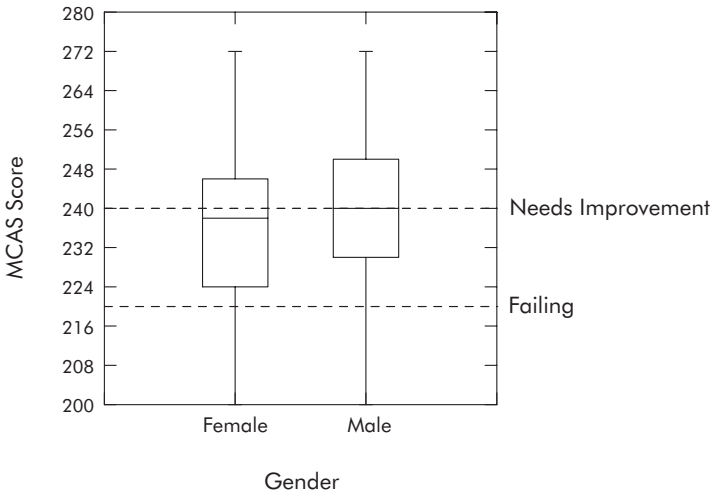

TABLE 4
Estimated "True Score" Correlations Among Grades and MCAS Scores in English, Math, and Science

	English Grade	*Math Grade*	*Science Grade*	*English MCAS*	*Math MCAS*	*Science MCAS*
English Grade	1.000					
Math Grade	0.583	1.000				
Science Grade	0.798	0.642	1.000			
English MCAS	0.593	0.373	0.568	1.000		
Math MCAS	0.557	0.536	0.628	0.701	1.000	
Science MCAS	0.485	0.354	0.537	0.688	0.763	1.000

below the cutoff score for "needs improvement" (Figure 6), while the median for boys falls approximately on the cutoff. The difference in proportion of girls and boys falling into the lowest two categories in science (55% of girls, 48% of boys) is much less dramatic than it is for math.

Correlations of Teacher-Assigned Grades and MCAS Scores

The HLM estimates of the "true score" correlations appear in Table 4. One useful way of examining these correlations is to look at the strength of association between teacher-assigned grades and MCAS scores within the same subject area. The estimated "true score" correlations between MCAS scores and teacher-assigned grades look fairly consistent across all the subjects, ranging from the low of .536 for the correlation of the math MCAS with math grades to a high of .593 for the estimated correlation of English MCAS with English grades. It is important to note that the estimated intercorrelations of the teacher-assigned grades in each subject and the intercorrelations of the MCAS scores in each subject are generally greater than the estimated correlations across the two assessments within a discipline. Teacher-assigned grades in English and teacher-assigned grades in science are estimated to correlate the most strongly at .798. The estimated "true score" correlation of math grades to English grades and math grades to science grades are .583 and .642, respectively. For the MCAS, the estimated "true score" correlation of math and science scores is the largest at .763. The estimated "true score" correlation between English MCAS scores and math MCAS scores is .701, and the estimated correlation between English MCAS scores and science MCAS scores is .688. The pattern of higher correlations within a method across disciplines suggests the possibility that

TABLE 5
*Estimated Regression Coefficients, Standard Errors, T-Ratios, and Probability Values from Hierarchical Linear Model Predicting Teacher-Assigned Grades and MCAS Scores in English, Math, and Science**

	Fixed Effect	*Coefficient*	*Standard Error*	*T-Ratio*	*P-Value*
English					
Grades	Intercept	3.088	0.761	4.058	< .0005
	African American	–2.978	1.959	–1.520	0.128
	Latino/a	–10.549	3.289	–3.207	0.002
	Asian	0.926	1.528	0.606	0.544
	"Other"	–4.844	2.663	–1.819	0.068
	Female	2.888	0.725	3.986	< .0005
	School 1	–4.488	0.972	–4.620	< .0005
	School 2	–2.283	1.058	–2.159	0.031
	School 4	–3.202	1.507	–2.125	0.033
	School 5	–11.152	2.315	–4.818	< .0005
	School 6	–9.727	1.171	–8.306	< .0005
Math					
Grades	Intercept	4.021	0.776	5.180	< .0005
	African American	–1.341	2.013	–0.666	0.505
	Latino/a	–2.744	3.371	–0.814	0.416
	Asian	1.324	1.557	0.850	0.395
	"Other"	–2.576	2.724	–0.945	0.345
	Female	1.672	0.740	2.258	0.024
	School 1	–2.279	0.990	–2.301	0.021
	School 2	–8.264	1.078	–7.665	< .0005
	School 4	–1.369	1.549	–0.884	0.377
	School 5	–2.577	2.415	–1.067	0.287
	School 6	–12.547	1.197	–10.482	< .0005
Science					
Grades	Intercept	5.606	0.767	7.314	< .0005
	African American	–3.720	1.967	–1.891	0.058
	Latino/a	–7.551	3.307	–2.283	0.022
	Asian	0.720	1.540	0.467	0.640
	"Other"	–9.642	2.679	–3.599	0.001
	Female	1.175	0.730	1.611	0.107
	School 1	–4.893	0.979	–4.998	< .0005
	School 2	–7.257	1.066	–6.811	< .0005
	School 4	–8.578	1.514	–5.664	< .0005
	School 5	–7.251	2.304	–3.147	0.002
	School 6	–11.321	1.179	–9.600	< .0005

* Because of the collinearity of student's race/ethnicity, school attended, and participation in the METCO program, an effect of METCO participation is not estimated in these models. In order to investigate whether the socioeconomic status or some other characteristic of the METCO participants might be enhancing or suppressing effects of race/ethnicity, we estimated a model in which the school dummies were dropped from the

TABLE 5 *(continued)*
Estimated Regression Coefficients, Standard Errors, T-Ratios, and Probability Values from Hierarchical Linear Model Predicting Teacher-Assigned Grades and MCAS Scores in English, Math, and Science

	Fixed Effect	*Coefficient*	*Standard Error*	*T-Ratio*	*P-Value*
English					
MCAS	Intercept	3.655	0.716	5.106	< .0005
	African American	–2.070	1.739	–1.190	0.234
	Latino/a	–8.069	2.976	–2.711	0.007
	Asian	–1.444	1.439	–1.004	0.316
	"Other"	–2.105	2.442	–0.862	0.389
	Female	2.727	0.675	4.039	< .0005
	School 1	–2.844	0.914	–3.114	0.002
	School 2	–4.364	1.001	–4.359	< .0005
	School 4	–6.191	1.357	–4.563	< .0005
	School 5	–9.133	1.836	–4.973	< .0005
	School 6	–7.182	1.087	–6.607	< .0005
Math					
MCAS	Intercept	7.682	0.733	10.485	< .0005
	African American	–5.880	1.780	–3.302	0.001
	Latino/a	–7.862	3.048	–2.579	0.010
	Asian	–0.839	1.473	–0.569	0.569
	"Other"	–4.344	2.500	–1.737	0.082
	Female	–1.899	0.691	–2.749	0.006
	School 1	–3.891	0.935	–4.160	< .0005
	School 2	–8.505	1.025	–8.299	< .0005
	School 4	–11.134	1.393	–7.990	< .0005
	School 5	–12.453	1.817	–6.853	< .0005
	School 6	–10.252	1.115	–9.193	< .0005
Science					
MCAS	Intercept	4.846	0.770	6.296	< .0005
	African American	–5.635	1.870	–3.014	0.003
	Latino/a	–8.446	3.198	–2.641	0.009
	Asian	–1.134	1.546	–0.734	0.463
	"Other"	–6.128	2.624	–2.335	0.020
	Female	–1.973	0.726	–2.717	0.007
	School 1	0.691	0.982	0.704	0.482
	School 2	–4.783	1.077	–4.443	< .0005
	School 4	–7.575	1.468	–5.162	< .0005
	School 5	–6.151	1.959	–3.139	0.002
	School 6	–6.750	1.171	–5.766	< .0005

analysis and a dummy variable for METCO participation was included. There was no systematic change in the values of the race/ethnicity coefficients; some grew larger, others smaller. Importantly, there was no change in which race/ethnicity effects were significantly different from zero, and of the six outcomes, METCO was only related to teacher-assigned grades in English ($p = .032$), where participation was negatively related to the grades.

the assessment method may play a significant role in how students perform on these two types of assessments.

Effects of Student Characteristics on Teacher-Assigned Grades

There are numerous significant differences between students in their teacher-assigned grades that are associated with their race and gender. Coefficients estimated by our two-level hierarchical model for teacher-assigned grades are shown in Table 5.

While in all three disciplines the coefficients estimated for the difference between African American and White students' grades are negative — indicating that, on average, African American students receive lower grades in these subjects than White students — none of these coefficients is significantly different from zero at an alpha level of 0.05. In the case of science grades, the p-value of 0.058 is very close to the accepted level of 0.05 and worthy of consideration, given the small number of African American students in the sample.

For all three subjects, the coefficients comparing the average grades of Latino/Latina students to those of White students are negative. For both English and science grades, the estimated coefficients are statistically significant at the .05 level, meaning that Latino/Latina students are estimated to have lower grades than White students. In particular, the coefficient representing the difference in grades for Latino/Latina versus White students in English (–10.549), about one full standard deviation, is the largest of all the coefficients for teacher-assigned grades and is considered to be a "large" effect (Cohen, 1988). The coefficient for grades in math is not significantly different from zero.

All the estimated coefficients for the difference between Asian students' and White students' grades are positive and fairly small. Asian students, on average, earn higher grades than White students in all three subjects, but none of these coefficients was significantly different from zero.

The students categorized in the "other" race/ethnicity category look somewhat similar to the Latino/Latina students. In science, students in the "other" category receive significantly lower grades than White students. The difference between "other" and White students' grades in English is just outside the accepted 0.05 level of significance ($p = .068$). In science, students grouped in the "other" category have predicted grades even lower than those for Latino/Latina students. The difference in math grades for the "other" students is about the same as that for Latino/Latina students and, similarly, it does not meet a standard level of statistical significance.

For all three subjects, the estimated coefficients for female students show girls receiving higher teacher-assigned grades than boys. The male-female differences in math and English grades are somewhat smaller than the race/ethnicity differences but are statistically different from zero. The coefficient

for the male-female difference in science grades is small and not statistically different from zero.

Effects of Student Characteristics on MCAS Scores

In general, the effects of race/ethnicity and gender on MCAS scores are somewhat similar to the effects on grades, but the pattern and especially the size of the estimated coefficients are not identical to the pattern and effect sizes seen in the coefficients for grades. The differences between the two groups of coefficients will be discussed in greater depth in the next section of our results, which explores differences in equitability between teacher-assigned grades and MCAS scores.

Differences in MCAS scores were found among students of different races/ethnicities. African American students, on average, score significantly lower than White students on the math and science portions of the MCAS, with a medium effect greater than half a standard deviation (Cohen, 1988). While the coefficient for the African American versus White difference on the English portion of the test is also negative, it is not significantly different from zero. Latino/Latina students have estimated MCAS scores significantly below those of White students in all three disciplines. Asian students could not be distinguished statistically from White students in their MCAS scores, although all three effects were negative. Students characterized as "other" have predicted MCAS scores significantly below those of White students in science. The coefficient for the difference in math scores is slightly less than half a standard deviation, but does not meet the conventional 0.05 level of statistical significance, while there is no statistically significant difference between "other" students and White students in English.

Gender differences in MCAS scores were also found. Girls score significantly lower, on average, than boys in both the science and math portions of the MCAS, but they have significantly higher scores on the English portion of the test. The male-female differences fall within the range of "small" effect sizes (Cohen, 1988), but are clearly statistically significant from zero because each gender is about one-half of the sample.

Between School Differences in MCAS Scores

Fixed effects estimates were included in these models to account for average differences between schools. The fixed effects estimates in the teacher-assigned grades level-2 model are relatively meaningless because grading in each school is idiosyncratic and really only has meaning within a school. Meanwhile, the fixed effects estimates for differences in school means on the three sections of the MCAS are of some interest because the MCAS tests are scored using the same system statewide. The fixed effects estimates for the schools are interpreted as differences from the reference school/dis-

trict. School/District 3 was selected as the reference school because it had the largest number of students among the schools in the sample. The other five schools scored significantly lower than School/District 3 in all three of the subject areas, with only one exception — School/District 1 had a slightly higher average score in the science portion of the MCAS than School/District 3, but the difference was not statistically significant. The three schools in District 4 (Schools 4, 5, and 6) consistently had the lowest MCAS results.

Equitability Differences between Teacher-Assigned Grades and MCAS Scores

To shed empirical light on the question of whether differences in equitability might arise if MCAS scores were to supersede teacher-assigned grades as a basis of academic decisionmaking, we conducted a series of hypothesis tests to see whether the effects of race/ethnicity and gender on teacher-assigned grades are different from the effects of race/ethnicity and gender on MCAS scores. In other words, applying our definition of equitability, is the MCAS more or less equitable than teacher-assigned grades? Would using the MCAS to make key academic decisions such as promotion or placement in advanced courses result in greater educational inequality than relying on grades?

African American students achieve near parity with White students when teachers assign grades in math, but on the MCAS they score lower. The estimated coefficients predicting teacher-assigned math grades and MCAS scores for African Americans are significantly different from one another ($\chi^2 = 5.595$, $d.f. = 1$, $p = .017$). In English, African Americans, on average, receive lower grades than Whites and score lower on the MCAS. The gap between the races is fairly similar for both assessments, and, indeed, the coefficient estimates are not statistically different ($\chi^2 = 0.265$, $d.f. = 1$, $p > .500$). African American students, on average, score lower than White students on the science section of the MCAS, and they receive lower grades as well. Although the MCAS coefficient estimate is somewhat larger than the estimate for grades, the two coefficients are not significantly different ($\chi^2 = 0.996$, $d.f. = 1$, $p > .500$). Thus, only the math portion of the MCAS appears to worsen the relative position of African Americans.

Latino/Latina students, on average, are predicted to have lower teacher-assigned grades and MCAS scores across all three subjects than White students. While Latino/Latina students are predicted to be assigned slightly lower grades than White students, on average, they fare worse than White students on the MCAS. Although the gap in MCAS math scores appears larger than the gap in math grades, this difference does not achieve conventional statistical significance ($\chi^2 = 2.549$, $d.f. = 1$, $p = .106$), but, given the limited statistical power of the study, may be worthy of consideration. In English, Latino/Latina students are predicted to receive considerably lower grades and to score lower on the MCAS than White students. This gap seems fairly similar for both assessments, and the two coefficients are not statistically different from one another ($\chi^2 = 0.709$, $d.f. = 1$, $p > .500$). The gaps be-

tween Latino/Latina and White students in science look very much like the gaps in English and, once again, the gaps in teacher-assigned grades and MCAS scores cannot be distinguished from one another statistically (χ^2 = 0.077, *d. f.* = 1, $p > .500$). Thus, with the possible exception of English, the MCAS does not appear to increase the inequities for Latino/Latinas.

Asian students on average earn higher grades than White students in all three subjects, but they score lower than Whites on the three sections of the MCAS, though none of the differences is significantly different from zero. Although the effects have opposite signs, none is significantly different from one another (math: $\chi^2 = 2.153$, *d. f.* = 1, $p = .138$; English ($\chi^2 = 3.055$, *d. f.* = 1, $p = .077$; science: $\chi^2 = 1.535$, *d. f.* = 1, $p = .213$). Asian students appear to fare about the same on both assessments.

Students grouped in the "other" category for race/ethnicity on average receive lower grades from their teachers and score lower on the MCAS than Whites. The differences between the coefficients for this group of students are not significantly different from zero (English: $\chi^2 = 1.327$, *d. f.* = 1, $p = .248$; math: $\chi^2 = 0.467$, *d. f.* = 1, $p > .500$; science: $\chi^2 = 1.817$, *d. f.* = 1, $p = .174$).

While girls tended to receive slightly higher grades in math than boys on average they scored somewhat lower than boys on the math portion of the MCAS. The two estimated coefficients for girls' assigned grades and MCAS scores were significantly different from one another ($\chi^2 = 25.723$, *d. f.* = 1, $p < .0005$); in other words, girls do significantly worse on the math MCAS than on teacher-assigned grades. In science, the pattern seen with math grades and MCAS scores is repeated. Science grades for girls are somewhat higher on average than those for boys, although the difference was not found to be significantly different from zero. On the other hand, girls score significantly lower on the science portion of the MCAS than boys ($\chi^2 = 19.444$, *d. f.* = 1, $p < .0005$). For these two subjects, math and science, it is difficult to say whether the tests or grades are more inequitable without any further indicators of the "true" abilities of girls and boys. In English, girls are predicted to do better than boys on both teacher-assigned grades and MCAS scores. In this subject, the coefficients predicting the gender gaps look similar across the two measures, and they are not significantly different from one another ($\chi^2 = 0.062$, *d. f.* = 1, $p > .500$).

Results Compared to National Assessment of Educational Progress (NAEP)

Because we had a limited sample in both size and diversity of the schools and the students in them, we elected to compare our results for the equitability of the MCAS to results seen in the National Assessment of Educational Progress (NAEP). The NAEP results will be more representative of all students attending school in Massachusetts, in contrast to our sample, which provides information on students attending only suburban schools. Although we believe that the comparison of inequitability in our sample to inequitability in the NAEP sample is a useful one, there are some important distinctions that

may affect the interpretation of the results. The most important distinction is that, unlike the use of a fixed effects model to eliminate differences in school mean scores on the MCAS, the NAEP results make no such accommodations. Because NAEP assesses the performance of a representative sample of students, and because each student answers only a sample of questions from a larger content domain, NAEP results allow us to make valid generalizations about the performance of all students in a particular academic subject. Although NAEP does not generate student-, school-, and district-level data, it breaks down scores for White and minority students as well as males and females. In our comparisons, we used results from the most recent administration of the NAEP in Massachusetts. Since NAEP and MCAS scores are not scored in the same way, we calculated effect sizes, which express test scores in standard deviation units (SD). Cohen (1988) has provided guidelines for defining "small," "medium," and "large" effects, which correspond to effect sizes of approximately .20, .50, and .80.

Table 6 compares effect sizes based on the 1998 MCAS with test scores from the most recent administrations of the reading (1998), mathematics (1996), and science (1996) NAEP to eighth graders in Massachusetts. The NAEP scores reflect the performance of a representative sample of students and schools across suburban, urban, and rural areas in Massachusetts, whereas the MCAS scores reflect the performance of a nonrepresentative sample of test-takers drawn from four suburban districts in metropolitan Boston.[5]

Expressed in terms of effect sizes, the African American–White gap on the NAEP is –0.79 SD in reading, –1.05 SD in math, and –1.34 SD in science. These large effect sizes are consistent with the African American–White test-score gap, which has ranged between 0.80 SD and 1.20 SD on a variety of national assessments (Hedges & Nowell, 1998). In our sample, however, the African American–White gap on the MCAS is considerably smaller, with effect sizes of –0.21 SD in English, –0.59 SD in math, and –0.56 SD in science. These differences are approximately 0.50 SD smaller for the MCAS than for the NAEP in reading and math, and nearly 0.75 SD smaller in science. There are three plausible explanations for this finding. First, because the NAEP results do not factor out differences among schools, this effect might be at least partly attributable to the phenomenon of more African American students attending lower performing schools than White students. The second possible explanation is closely related to the first, which is that our deliber-

[5] According to the U.S. Office of Management and Budget, "a metropolitan statistical area (MSA) is a geographic area consisting of a large population nucleus together with adjacent communities that have a high degree of economic and social integration with that nucleus" (Standards for Defining Metropolitan and Micropolitan Statistical Areas, 65 Fed. Reg. 2000). NAEP results are provided for students attending public schools located in an MSA central city, MSA urban fringe (non-central cities of MSA), and rural/small towns (non-MSAs). The Boston metropolitan area comprises five central cities; the remainder of the towns and counties, including the four school districts in our sample, are officially part of the "urban fringe" but are referred to as "suburbs" in this article.

TABLE 6
*Achievement Disparities by Gender and Race/Ethnicity on Grade-8 Massachusetts NAEP Reading (1998), Mathematics (1996), and Science (1996), and Grade-8 MCAS (1998) Reading, Mathematics, and Science**

	NAEP	MCAS	NAEP-MCAS
	Reading (1998)	*English*	
Male (vs. Female)	–0.35	–0.27	–0.08
African American (vs. White)	–0.79	–0.21	–0.58
Latino/Latina (vs. White)	–0.97	–0.81	–0.16
Asian (vs. White)	–0.24	–0.14	–0.10
	Mathematics (1996)	*Math*	
Male (vs. Female)	0.05	0.19	–0.14
African American (vs. White)	–1.05	–0.59	–0.46
Latino/Latina (vs. White)	–1.31	–0.79	–0.52
Asian (vs. White)	–0.20	–0.08	–0.12
	Science (1996)	*Science*	
Male (vs. Female)	0.17	0.22	–0.05
African American (vs. White)	–1.34	–0.56	–0.78
Latino/Latina (vs. White)	–1.31	–0.84	–0.47
Asian (vs. White)	–0.40	–0.11	–0.29

Source: NAEP Reading: Ballator & Jerry, 1998; NAEP Math: Reese, Jerry, & Ballator, 1997; NAEP Science: O'Sullivan, Jerry, Ballator, & Herr, 1997.

* Sample sizes for NAEP Reading (1998): male (n = 1060), female (n = 1081); African American (n = 142), White (n = 1642); Latino/Latina (n = 228), White (n = 1642); Asian (n = 114), White (n = 1642). NAEP Math (1996): male (n = 1150), female (n = 1130); African American (n = 150), White (n = 1833); Latino/Latina (n = 175), White (n = 1833); Asian (n = 102), White (n = 1833). NAEP Science (1996): male (n = 1159), female (n = 1128); African American (n = 153), White (n = 1840); Latino/Latina (n = 177), White (n = 1840); Asian (n = 91), White (n = 1840).

ate selection of suburban schools does, in fact, tend to standardize such things as the quality of the curriculum and the teaching standards. The third point, and one that future researchers may wish to pursue further, is that the curriculum-based MCAS may be more equitable than a test such as NAEP, which focuses on a broader range of skills and aptitudes.

A similar pattern exists for Latino/Latina students. By Cohen's guidelines, the Latino/Latina-White MCAS gap is consistently large across all three subjects (–0.81 SD reading, –0.79 SD math, –0.84 SD science). Compared to NAEP, the MCAS gap is smaller by a magnitude of 0.16 SD in reading, 0.52 SD in math, and 0.47 SD in science. So once again, in the case

of Latino/Latina students the same three plausible explanations showing a narrower gap between Whites and Latinos/Latinas on MCAS than on NAEP are worth considering.

Finally, the Asian-White gap is small to medium on the NAEP, depending on the subject. It ranges from –0.24 SD in reading, to –0.20 SD in math, to –0.40 SD in science, compared to –0.14 in reading, –0.08 in math, and –0.11 in science on the MCAS. Once again, the same three explanations hold for Asian students.

We observe small gender differences on the NAEP and MCAS across all three subjects. In the 1998 cohort, boys score an average of 0.35 SD lower than girls on NAEP reading and 0.27 SD lower on the MCAS English. Boys perform somewhat better than girls in math and science on the 1996 NAEP (0.05 SD math, 0.17 SD science). On the MCAS math and science, boys score approximately 0.20 SD higher than girls. Because the proportion of boys and girls in schools throughout the Commonwealth of Massachusetts will always be close to 50:50, explanations of gender differences would not include any theory based on the proposition that observed gender differences are the result of boys and girls attending largely different schools, perhaps the most viable explanation for racial/ethnic differences seen in the comparison of NAEP and MCAS scores. This leaves two plausible explanations for differences in the "gender gap" of NAEP and MCAS. The first is that there is something about the curriculum or teaching in the suburban schools in our sample that either increases (math and science) or decreases the gap (English). The other plausible explanation is that the gender equitability of the two tests is different. If the latter theory can be supported, it suggests that the math portion of the MCAS, which yields a four-times larger gap between boys and girls, should be scrutinized.

Limitations

Limiting a sample to suburban schools, in addition to limiting the generalizability of the study, poses a large threat in a study of racial/ethnic equitability. Most suburban schools are overwhelmingly White, which has consequences for use of such districts in a statistical analysis. Specifically, the ability to find a difference, particularly a small one, between two or more groups of students is dependent on the precision with which average scores for each of the groups can be estimated. Smaller sample sizes for a given group result in estimates with lower precision for that group. Even if average scores for a White majority may be estimated with great precision, the ability to detect a difference from another group will be limited by the precision with which the other group's score may be estimated. This is a question of statistical power, the ability to find a difference in a sample when one exists in the population. The lower the power, the less likely such a difference will

be detected in the sample, particularly if the difference in the population is a small one.

The implications for the present study are that any findings of differences between racial/ethnic groups and between boys and girls may be readily accepted as existing in the population, in this case the four districts using a conventional level of probability to reject the null hypothesis of no difference. However, failure to find a difference for racial/ethnic groups in the sample must be interpreted in the light of statistical power of the study. In particular, a finding of no differences between small groups or even between a small group and the largest group should not be considered definitive evidence of a lack of difference, but a possible consequence of low statistical power. For example, moderate-looking differences between African American and White students in teacher-assigned grades in all three disciplines do not achieve accepted statistical significance. The lack of a finding for these three effects may be the result of a lack of power to find effects of this magnitude, given the number of African American students in the sample. Because some of the estimated differences between Latino/Latina students and White students are larger than those for the African American–White gap, we find more statistically significant differences. Because the sample is roughly half female, fairly small differences between girls and boys achieve statistical significance.

Statistical power also limits our ability to find small differences within ethnic groups; that is, we cannot detect small differences in the equitability of the MCAS tests compared with the equitability of teacher-assigned grades. The resulting lack of statistical power may explain why, although some coefficients for African American–White student differences vary by the assessment method, the coefficients do not turn out to be statistically different from one another. Also, for Asian students, who have slightly higher grades but slightly lower scores on the MCAS than White students, the difference between these estimated coefficients, despite their opposite signs, does not achieve conventional statistical significance.

Due to limited statistical power for some of our comparisons, we choose not to emphasize our "nonfindings" (i.e., findings that did not achieve conventional statistical significance). Before concluding that MCAS scores do not appear to create large gaps in equitability in these instances, studies that deliberately include larger numbers of students in each of these racial/ethnic categories should be conducted. On the other hand, when we do find equitability differences in these data, one can conclude with a degree of statistical confidence that differences exist in the population. In other words, while the sample sizes for each of these groups limit our ability to claim that there are no differences in the equitability of the two assessments, they do not in any way limit the ability to make claims based on differences that we do find in the data.

Discussion

Do high-stakes testing programs worsen educational outcomes for racial/ethnic minorities and girls? In comparison to teacher-assigned grades, MCAS hurts the average competitive position of African American students in math and of girls in math and science. There is suggestive evidence in this limited sample that the MCAS may also have a differential impact on Latinos/Latinas in math. Racial/ethnic group comparisons of mathematics achievement suggest that the African American–White gap and Latino/Latina-White gap on MCAS scores is larger by about 0.50 standard deviations than differences on teacher-assigned grades. MCAS scores place girls at a disadvantage to boys in math and science, where the gender gap is larger by approximately one-third standard deviation than grades.

In arguing that the equitability gaps are maintained or increased under high-stakes testing, we explicitly do not assert that this demonstrates that high-stakes tests are more biased than grades or even biased at all. Bias would exist if we knew that one or both of the assessments resulted in an evaluation of students that differed from their true abilities, but, as Supovitz and Brennan (1997) noted, "to judge which assessment is closer to real student performance, we must know each child's true ability" (p. 496). Unfortunately, we have no additional evidence of the students' underlying abilities available to us. While it is tempting to conclude, for example, that when girls who get better grades in science and math score lower on a standardized test, the test itself or perhaps its methods are biased against girls, it is equally plausible that the teacher-assigned grades, which may incorporate nonacademic factors such as classroom behavior, may be biased against boys. Why, then, do high-stakes test scores sustain and, in some cases, expand the achievement gap?

Unlike standardized test scores, classroom grades represent student competency in a variety of cognitive and noncognitive domains. By assessing a broad range of skills, behaviors, and attitudes, teachers' strong evaluations of student effort and initiative may compensate for relatively poorer performance on academic tasks. Grading standards often rely more heavily on subjective criteria, including a teacher's attitudes toward minority students (Entwisle & Alexander, 1988; Haney, 1993; Leiter & Brown, 1985), personal educational philosophy (Brookhart, 1993; Cizek, Fitzgerald, & Rachor, 1996; Waltman & Frisbie, 1994), and perceptions of student effort and conduct (Cizek, 1996; Stiggins & Conklin, 1992). Teachers' concerns about equal opportunity and outcomes may also encourage grading policies that reward effort and initiative (Jencks, 1985). Since African American children appear to have a stronger desire to please teachers than White children (for a complete review, see Ferguson, 1998), and since teachers reward student initiative and effort, we would expect to find more equitable outcomes on school grades. Moreover, among African American children, teacher expectations appear to exert a stronger effect on grades than test scores (Entwisle & Alex-

ander, 1988; Jussim, Eccles, & Madon, 1996; McCandless, Roberts, & Starnes, 1972).

Similarly, teacher-assigned grades appear to favor Asians over Whites just slightly, whereas, after controlling for the school attended, the small advantage is reversed on the MCAS. In other words, the small positive advantage for Asians on grades becomes negative on the MCAS, although only the Asian-White gap in English ($p = .079$) is close to reaching the .05 level of significance. Therefore, our conclusion regarding Asian achievement is speculative and tentative. Our results, however, do not contradict research on Asian students' academic success, which suggests that Asian parents of all educational levels push their children to do well in school, as measured by teacher-assigned grades.

In sum, using either the MCAS or grades to assess student achievement would have an uneven effect on the competitive position of each respective racial/ethnic group, especially since performance varies across subjects and assessment formats.

Our findings show that boys in suburban schools outperform girls on the math and science MCAS, whereas the reverse is true for teacher-assigned grades. Two related bodies of research help to explain this discrepancy. First, social scientists (Gallagher, 1998; Hyde et al., 1990; Steinkamp & Maehr, 1984; Willingham & Cole, 1997) have suggested that standardized tests in quantitative subjects such as math and science may pose a greater challenge to girls than boys. Gallagher (1998) points out that, "in terms of the discrepancy between course grades and standardized test performance in mathematics, course work relies heavily on assessing and retrieval of information, skills at which girls tend to excel. Standardized tests, on the other hand, may rely more heavily on the quick mental manipulation tasks at which boys tend to excel" (p. 305). When student performance is assessed using grades, girls appear to enjoy several advantages over boys. While teachers believe that boys and girls are equally talented in math, they often evaluate girls as harder workers. Based on this finding, Jussim, Eccles, and Madon (1996) conclude: "Because high effort is generally viewed positively by teachers and others, and because teachers rewarded supposedly harder-working students with higher grades, this bias seems to favor girls" (p. 332). We acknowledge that such findings are speculative and tentative since we did not collect data on teacher attitudes toward boys and girls. Nevertheless, it is clear that the size and even the direction of the gender gap depends on the type of assessment used to evaluate achievement.

Even if standardized tests are not used in a gatekeeping role to determine who may move on to advanced classes, performance on standardized tests may have important implications for students' own long-term decisions regarding course selection and career choices (Hewitt & Seymour, 1991; Steele, 1999). According to Wilder and Powell's (1989) literature review of gender differences in test performance, "males and females may be affected

differently by their success (or lack thereof) as that success is reflected in test performance. Lesser performance on (say) measures of mathematical skill, whatever their origin, may cause girls to lower their aspirations, lose their self-confidence, take courses in areas other than the quantitative ones, and/or conclude that certain domains are the province of males" (p. 31).

Do our results suggest anything about the validity of either teacher-assigned grades or the MCAS as measures of achievement? If both high-stakes tests and teacher-assigned grades were valid and reliable measures of student mastery of subject matter, the correlations between the two measures within a subject would be very strong. In this instance, however, even after correcting for measurement error (i.e., unreliability), the correlations are only modest, about .5 to .6, a figure that is in line with the reported correlation between teacher marks and standardized tests in most other studies (Linn, 1982). We must conclude, based on this evidence, that one or both of these assessments are imperfect measures of true achievement. The strong correlations of teacher-assigned grades across the subjects, and thus across teachers, is striking and suggests that qualities such as effort, work ethic, and ability to meet teachers' expectations play a significant role in assigning grades. While not definitive, the evidence found in these "true score" correlations raises the question of whether the assessment method itself, either teacher-assigned grades or standardized tests, may play a large role in determining a student's score. If this is true, then qualities considered to be separate from underlying knowledge of the subject matter, such as test-taking savvy, ability to work under a tight time limit, familiarity with the exact format of the assessment, or a pattern of behavior that is generally more pleasing to teachers, may be playing a large role in determining MCAS scores and/or grades. To the extent that these qualities and not knowledge and ability in the subject may be contributing to the widening of some of the equitability gaps, the effect of the assessment format of high-stakes tests merits further scrutiny.

Taken together, equitability disparities on the MCAS underscore the need for extreme caution in the implementation of high-stakes testing programs. Indeed, many legal safeguards and policy prescriptions have been proposed to ensure that high-stakes tests like the MCAS do not perpetuate inequities among students based on race/ethnicity and gender. Two proposals in particular seem especially promising because each conforms to professional norms governing educational testing and each attempts to provide contextual information for interpreting a single test score. First, high-stakes test scores, if used appropriately, should drive substantive improvements in curriculum and instruction by helping educators design instructional programs that raise achievement (American Educational Research Association et al., 1999; Elmore & Rothman, 1999; Heubert & Hauser, 1998; Popham, 1987).

Second, educators should not rely on standardized tests as the sole criterion for making high-stakes decisions regarding individual students, and

should incorporate other academic factors to supplement test scores (Coleman, 1998; Darling-Hammond & Falk, 1997; Glaser, 1990; Heubert & Hauser, 1998). An accountability system based on test scores and teacher marks should capitalize on the strengths of both measures. Standardized test scores, by definition, supply the general public with a common yardstick for measuring student achievement across a diversity of educational settings. Grading standards, on the other hand, may vary across classrooms, schools, districts, and regions of the country, thus failing to provide a consistent and objective picture of student achievement (Murnane & Levy, 1996; Office of Educational Research and Improvement, 1994; Puma et al., 1997). In short, standardized test scores produce more reliable and consistent measures of achievement than teacher-assigned grades. The disadvantage of standardized test scores, however, stems from their inherent limitation in providing a valid inference of student performance in a given academic subject.

Because questions on standardized tests such as the MCAS represent only a small sample of items from a broad academic domain, high-stakes test-score gains should be validated by other assessments. For example, the combination of test scores and grades might allow for a more valid assessment of achievement. Compared to a single high-stakes test, grades represent student performance over an entire school year on a variety of classroom assignments, which more comprehensively assess the domain of knowledge and skills that comprise each academic discipline.

Furthermore, teacher-assigned grades usually produce more equitable achievement results than standardized tests. As a result, some educational researchers (Haney, 2000; Heubert & Hauser, 1998) have argued that a sliding scale, involving grades and test scores, is needed to build a technically sound and equitable accountability system. A sliding scale, or compensatory model, would combine test scores and grades in making high-stakes decisions and allow higher grades to compensate for lower test scores. On the high-stakes TAAS examination, such a system would improve the competitive position of African American and Latino/Latina students (Haney, 2000). Combining tests and grades would also cushion the blow of a single failing test score. Gamoran (2000), in advocating a similar approach, recommends policies that "allow students to graduate on the basis of performance in courses and to use test performance at high school as an indicator of 'qualifications,' or mastery of specific curricular material" (p. 124).

When grades, test scores, and possibly some other achievement and non-achievement data are used to evaluate students and schools, incentives exist for educators to employ educational practices that develop cognitive mastery over a broad domain of academic knowledge and skills (Elmore & Rothman, 1999; Koretz, 1996; Linn, 2000). An accountability system based on multiple achievement indicators might also command strong political support because the public embraces the notion that schools should develop noncognitive outcomes such as good citizenship (Rothstein, 2000).

We believe that relying exclusively or heavily on high-stakes tests, such as the MCAS, to make critical academic decisions, such as the granting of a high school diploma, might dramatically set back girls and students of color. This alone should raise a call for increased scrutiny of the assessments themselves. To the extent that high-stakes tests may increase accountability in underperforming schools, typically those serving large minority populations, there is potential to improve the academic environment for groups now poorly served by the educational system. However, if the tests become a de jure standard for graduation or promotion with little reform of the school system, groups already marginalized may be further punished and potentially cut off disproportionately from further academic opportunities and employment.

While it is often difficult for researchers to gain access to data on individual students that include both teacher-assigned grades and scores on high-stakes test, as well as identifiers for gender, race/ethnicity, and socioeconomic status, such analyses offer perhaps the one best hope of anticipating how the increasing use of such tests may affect groups of students already behind in educational outcomes. While we choose to examine this question in a limited sample of suburban schools, we believe further research should be carried out in schools of all types, particularly urban and inner-city schools, where a disproportion of racial/ethnic minority students are educated. We also urge that during the development of high-stakes tests, all possible attention be given to the question of how various groups may be affected by choices made during that development.

References

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

Asher, H. B. (1983). *Causal modeling* (2nd. ed.). Thousand Oaks, CA: Sage.

Ballator, N., & Jerry, L. (1998). *The NAEP reading state report for Massachusetts*. Washington, DC: National Center for Education Statistics.

Barnett, R. C., Marshall, N. L., Raudenbush, S. W., & Brennan, R. T. (1993). Gender and the relationship between job experiences and psychological distress: A study of dual-earner couples. *Journal of Personality and Social Psychology, 64*, 794–806.

Barth, P., Brennan, J., Haycock, K., Mora, K., Ruiz, P., & Wilkins, A. (1998). *Education watch: The 1998 Education Trust state and national data book* (vol. 2). Washington, DC: Education Trust.

Benning, V., & Mathews, J. (1999, August 14). State tests fail 93% of schools in Virginia. *Washington Post*, pp. A1, A6.

Bishop, J. H., Moriarty, J. Y., & Mane, F. (1997). *Diplomas for learning, not seat time: The impacts of New York Regents examinations* (Working Paper 97-31). Ithaca, NY: Cornell University, School of Industrial and Labor Relations.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley.

Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement, 30,* 123–142.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment.* Thousand Oaks, CA: Sage.
Catterall, J. S. (1989). Standards and school dropouts: A national study of tests required for high school graduation. *American Journal of Education, 98*, 1–34.
Chubb, J. E., & Moe, T. M. (1990). *Politics, markets, and America's schools.* Washington, DC: Brookings Institution Press.
Cizek, G. J. (1996). Grades: The final frontier in assessment reform. *NASSP Bulletin, 80,* 103–110.
Cizek, G. J., Fitzgerald, S. M., & Rachor, R. E. (1996). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational Assessment, 3,* 159–179.
Clinton, W., & Gore, A. (1992). *Putting people first: How we can all change America.* New York: Times Books.
Cohen, D. K. (1996). Standards-based school reform: Policy, practice, and performance. In H. Ladd (Ed.), *Holding schools accountable* (pp. 99–127). Washington, DC: Brookings Institution Press.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
Coleman, A. (1998). Excellence and equity in education: High standards for high-stakes tests. *Virginia Journal of Social Policy and the Law, 6,* 81–112.
College Board. (1999). *Reaching the top: A report of the national task force on minority high achievement.* New York: Author.
Darling-Hammond, L., & Falk, B. (1997). Supporting teaching and learning for all students: Policies for authentic assessment systems. In A. L. Goodwin (Ed.), *Assessment for equitability and inclusion: Embracing all our children* (pp. 51–76). New York: Routledge.
Darling-Hammond, L., & Wise, A. E. (1985). Beyond standardization: State standards and school improvement. *Elementary School Journal, 85,* 315–336.
Eaton, S. E. (2001). *Blurring the race boundary: Black adults raised in urban neighborhoods and schooled in White suburbia.* New Haven, CT: Yale University Press.
Editors. (1994). Introduction (Symposium: Equity and Educational Assessment). *Harvard Educational Review, 64,* 1–3.
Elmore, R., Abelman, C., & Fuhrman, S. (1996). The new accountability in state education reform: From process to performance. In H. Ladd (Ed.), *Holding schools accountable* (pp. 65–98). Washington, DC: Brookings Institution Press.
Elmore, R. F., & Rothman, R. (Eds.). (1999). *Testing, teaching, and learning.* Washington, DC: National Academy Press.
Entwisle, D. R., & Alexander, K. L. (1988). Factors affecting achievement test scores and marks of Black and White first-graders. *Elementary School Journal, 88,* 449–471.
Ferguson, R. F. (1998). Teachers' perceptions and expectations and the Black-White test score gap. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 273–317). Washington, DC: Brookings Institution Press.
Ferguson, R. F. (1999). *Racial test-score trends 1971–1996: Popular culture and community academic standards* (Working Paper). Cambridge, MA: John F. Kennedy School of Government, Malcolm Wiener Center for Social Policy.
Figueroa, R. A., & Hernandez, S. (2000). *Testing Hispanic students in the United States: Technical and policy issues.* Washington, DC: U.S. Department of Education, President's Advisory Commission on Educational Excellence for Hispanic Americans.
Fletcher, M. A. (1998, October 23). A good-school, bad-grade mystery. *Washington Post,* pp. A1, A10.
Fuchs, L. S., Fuchs, D., Eaton, S., Hamlett, C., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children, 67,* 67–81.

Gallagher, A. (1998). Gender and antecedents of performance in mathematics testing. *Teachers College Record, 100*, 297–314.

Gamoran, A. (2000). High standards: A strategy for equalizing opportunities to learn? In R. D. Kahlenberg (Ed.), *A nation at risk* (pp. 93–126). New York: Century Foundation Press.

Glaser, R. (1990). *Testing and assessment, O tempora! O mores!* Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Goertz, M., Duffy, M., & Carlson-LeFloch, K. (2000). *State assessment and accountability systems: 50 state profiles, 1999-2000*. Philadelphia: Consortium for Policy Research in Education.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Oxford University Press.

Goldstein, H. (1995). *Multilevel statistical models (*2nd. ed.). New York: Halstead.

Gordon, S. P., & Reese, M. (1997). High stakes testing: Worth the price? *Journal of School Leadership, 7*, 345–368.

Grissmer, D., & Flanagan, A. (1998). *Exploring rapid achievement gains in North Carolina and Texas*. Washington, DC: National Educational Goals Panel.

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What do NAEP test scores tell us?* Santa Monica, CA: RAND Corporation.

Haertel, E. (1989). Student achievement tests as tools of educational policy: Practices and consequences. In B. R. Gifford (Ed.), *Test policy and test performance: Education, language, and culture* (pp. 25–50). Boston: Kluwer Academic.

Haney, W. (1993). Testing and minorities. In L. Weiss & M. Fine (Eds.), *Beyond silence: Class, race, and gender in United States schools* (pp. 45–73). Albany: State University of New York Press.

Haney, W. (2000). The myth of the Texas education miracle. *Education Policy Analysis Archives, 8*. Available on-line: http://epaa.asu.edu/epaa/v8n41/part6.htm

Hanushek, E. A., & Jackson, J. E. (1977). *Statistical methods for social scientists*. New York: Academic Press.

Haycock, K. (1996). *Education watch: The 1996 Education Trust state and national data book*. Washington, DC: Education Trust.

Hedges, L. V., & Nowell, A. (1998). Black-White test score convergence since 1965. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 149–181). Washington, DC: Brookings Institution Press.

Hess, F. (1999). *Spinning wheels: The politics of urban school reform*. Washington, DC: Brookings Institution Press.

Heubert, J., & Hauser, R., (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Hewitt, N. M., & Seymour, E. (1991). *Factors contributing to high attrition rates among science and engineering undergraduate majors*. Unpublished report to the Alfred P. Sloan Foundation.

Hirsch, E. D. (1996). *The schools we need and why we don't have them*. New York: Doubleday.

Hochschild, J., & Scott, B. (1998). The polls—trends: Governance and reformed public education in the United States. *Public Opinion Quarterly, 62*, 79–120.

Hoffman, J. V., Assaf, L., Pennington, J., & Paris, S. G. (in press). High stakes testing in reading: Today in Texas, tomorrow? *Reading Teacher*.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*, 139–155.

Jackson, K. T. (1985). *Crabgrass frontier*. New York: Oxford University Press.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25*, 285–306.

Jencks, C. (1985). Whom must we treat equally for educational opportunity to be equal? *Ethics, 98*, 518–533.

Jencks, C., & Phillips, M. (1999). Aptitude or achievement: Why do test scores predict educational attainment and earnings? In S. E. Mayer & P. E. Peterson (Eds.), *Earning and learning* (pp. 15–47). Washington, DC: Brookings Institution Press.

Johnson, J., & Immerwahr, J. (1994). *First things first: What Americans expect from the public schools: A report from Public Agenda*. New York: Public Agenda.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd. ed.). Chicago, IL: SPSS.

Jussim, L., Eccles, J., & Madon, S. (1996). Social perceptions, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. *Advances in Experimental Social Psychology, 28,* 281–387.

Kahlenberg, R. D. (2000). *All together now: The case for economic integration of the public schools*. Washington, DC: Brookings Institution Press.

Kain, J. F., & O'Brien, D. M. (2000). *Black suburbanization in Texas metropolitan areas and its impact on student achievement* (Working Paper). Dallas: University of Texas-Dallas, Center for the Study of Science and Society.

Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological Bulletin, 105,* 198–214.

Kingdon, J. W. (1995). *Agendas, alternatives, and public policies*. New York: Harper Collins College.

Klein, R. (2000, November 3). MCAS criticism rising in suburbs contrast with view from urban areas. *Boston Globe,* p. A1.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives, 8,* 1–26. Available on-line: http://epaa.asu.edu/epaa/v8n49/

Kleinert, H., Haig, J., Kearns, J. F., & Kennedy, S. (2000). Alternate assessments: Lessons learned and the roads to be taken. *Exceptional Children, 67,* 51–66.

Kohn, A. (2000). *The case against standardized tests: Raise the test scores, ruin the schools*. Portsmouth, NH: Heinemann.

Koretz, D. (1988). Arriving at Lake Wobegon: Are standardized tests exaggerating Achievement and distorting instruction? *American Educator, 12,* 8–15, 46–52.

Koretz, D. (1996). Value-added indicators of school performance. In E. Hanushek & D. W. Jorgensen (Eds.), *Improving America's schools: The role of incentives* (pp. 171–195). Washington, DC: National Academy Press.

Koretz, D., & Barron, S. I. (1998). *The validity of gains on the Kentucky Instructional Results Information System (KIRIS)*. Santa Monica, CA: RAND.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April 5). *The effects of high-stakes testing on achievement: Preliminary findings about generalizations across tests.* Paper presented at the annual meeting of the American Education Research Association and the National Council on Measurement in Education, Chicago.

Kornhaber, M., Orfield, G., & Kurlaender, M. (2001). *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: Century Foundation Press.

Leiter, J., & Brown, J. S. (1985). Determinants of elementary school grading. *Sociology of Education, 58,* 166–180.

Levine, D. U., & Eubanks, E. E. (1990). Achievement disparities between minority and nonminority students in suburban schools. *Journal of Negro Education, 59,* 186–194.

Linn, R. L. (1982). Ability testing: Individual differences, prediction, and differential prediction. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies* (pp. 335–388). Washington, DC: National Academy Press.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29,* 4–15.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average." *Educational Measurement: Issues and Practice, 9,* 5–14.

Long, J. S. (1983). *Covariance structure models: An introduction to LISREL*. Thousand Oaks, CA: Sage.

Madaus, G. F., & Clarke, M. (2001). The adverse impact of high-stakes testing on minority students: Evidence from one hundred years of test data. In M. Kornhaber, G. Orfield, & M. Kurlaender (Eds.), *Raising standards of raising barriers? Inequality and high-stakes testing in public education* (pp. 85–106). New York: Century Foundation Press.

Massachusetts Department of Education. (1998). *Guide to the Massachusetts Comprehensive Assessment System.* Malden, MA: Author.

Massachusetts Department of Education. (1999). *MCAS technical advisory report summary.* Malden, MA: Author.

Massey, D. M., & Denton, N. A. (1993). *American apartheid: Segregation and the making of the underclass.* Cambridge, MA: Harvard University Press.

Mathews, J. (1999, May 13). An analysis of Va. scores show ethnicity gaps. *Washington Post*, p. B1.

Mathews, J., & Benning, V. (2000, September 11). Va. voters negative on SOLS, polls says; 51% in survey believe state tests don't work. *Washington Post,* p. B1.

McNeil, L. M. (2000). Creating new inequalities: Contradictions of reform. *Phi Delta Kappan, 81*, 728–735.

McNeil, L. M., & Valenzuela, A. (2001). The harmful impact of the TAAS system of testing in Texas: Beneath the accountability rhetoric. In M. Kornhaber, G. Orfield, & M. Kurlaender (Eds.), *Raising standards of raising barriers? Inequality and high-stakes testing in public education* (pp. 127–150). New York: Century Foundation Press.

McCandless, B. B., Roberts, A., & Starnes, T. (1972). Teachers' marks, achievement scores, and aptitude relations with respect to class, race, and sex. *Journal of Educational Psychology, 63*, 153–159.

Miller, L. S. (1999). Promoting high academic achievement among non-Asian minorities. In E. Y. Lowe (Ed.), *Promise and dilemma: Perspectives on racial diversity and higher education* (pp. 47-91). Princeton, NJ: Princeton University Press.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*, 707–726.

Murnane, R. J., & Levy, F. (1996). *Teaching the new basic skills.* New York: Free Press.

Muthén, B. (2001). Latent variable mixture modeling. In G A. Marcoulides & R. E. Schumacher (Eds.), *New developments and techniques in structural equation modeling* (pp. 1–33). Hilldale, NJ: Lawrence Erlbaum Accociates.

National Center for Education Statistics. (1997). *Reading and mathematics achievement: Growth in high school.* Washington, DC: Author.

National Governor's Association. (1986). *Time for results.* Washington, DC: Author.

O'Day, J., & Smith, M. S. (1993). Systemic reform and educational opportunity. In S. Fuhrman (Ed.), *Designing coherent educational policy* (pp. 250–312). San Francisco: Jossey-Bass.

Office of Educational Research and Improvement. (1994). *What do student grades mean? Differences across schools.* Washington, DC: United States Department of Education, Office of Educational Research and Improvement.

Ogbu, J. (1994). Racial stratification and education in the United States: Why inequality persists. *Teachers College Record, 96*, 264–298.

Orfield, G. (1996). The growth of segregation. In G. Orfield & S. Eaton (Eds.), *Dismantling desegregation* (pp. 53–72). New York: New Press.

Orfield, G., & Ashkinaze, C. (1991). *The closing door.* Chicago: University of Chicago Press.

O'Sullivan, C. Y., Jerry, L., Ballator, N., & Herr, F. (1997). *NAEP 1996 science state report for Massachusetts.* Washington, DC: National Center for Education Statistics.

Phelps, R. (1998). The demand for standardized student test. *Educational Measurement: Issues and Practice, 17*, 5–23.

Phelps, R. (1999). *Why testing experts hate testing.* Washington, DC: Thomas B. Fordham Foundation.

Phillips, M., Crouse, J., & Ralph, J. (1998). Does the Black-White test score gap widen after children enter school? In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 229–272). Washington, DC: Brookings Institution Press.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan, 68*, 679–682.

Prell, J. M., & Prell, P. A. (1986). *Improving test scores—teaching test-wiseness: A review of the literature.* Bloomington, IN: Phi Delta Kappa, Center on Evaluation, Development, and Research.

Public Agenda. (2000). *Survey finds little sign of backlash against academic standards or standardized tests.* New York: Author. Available on-line: http://www.publicagenda.org/aboutpa/pdf/standards-backlash.pdf

Puma, M. J., Karweit, N., Price, C., Ricciuti, A., Thompson, W., & Vaden-Kiernan, M. (1997). *Prospects: Final report on student outcomes.* Cambridge, MA: Abt Associates.

Quality counts 1999: Rewarding results, punishing failure. (1999, January 11). Bethesda, MD: Education Week,

Raudenbush, S. W., Brennan, R. T., & Barnett, R. C. (1995). A multivariate hierarchical linear model for studying psychological change within married couples. *Journal of Family Psychology, 9*, 161–174.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2000). *HLM5: Hierarchical linear and nonlinear modeling.* Chicago: Scientific Software.

Ravitch, D. (1995). *National standards in American education.* Washington, DC: Brookings Institution Press.

Reardon, S. F. (1996, April 8). *Eighth grade minimum competency testing and early high school dropout patterns.* Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Reese, C. M., Jerry, L., & Ballator, N. (1997). *NAEP 1996 Mathematics state report for Massachusetts.* Washington, DC: National Center for Education Statistics.

Richard, A. (2000, October 18). Remodeling suburbia. *Education Week, 20*, 29–30.

Rothstein, R. (1998). *The way we were?* New York: Century Foundation Press.

Rothstein, R. (2000). Toward a composite index of school performance. *Elementary School Journal, 100*, 409–441.

Sacks, P. (1999). *Standardized minds: The high price of America's testing culture and what we can do to change it.* Cambridge, MA: Perseus Books.

Smith, M. L., Heinecke, W., & Noble, A. J. (1999). Assessment policy and political spectacle. *Teachers College Record, 101*, 157–191.

Smith, M. S., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing* (pp. 233–267). New York: Falmer Press.

Spencer, J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 34*, 4–28.

Stage, E. K., Kreinberg, N., Eccles, J., & Becker, J. R. (1985). Increasing the participation and achievement of girls and women in mathematics, science, and engineering. In S. Klein (Ed.), *Handbook for achieving sex equity through education* (pp. 237–268). Baltimore: Johns Hopkins University Press.

Steele, C. M. (1999). A threat in the air: How stereotypes shape intellectual identity and performance. In E. Lowe (Ed.), *Promise and dilemma* (pp. 92–128). Princeton, NJ: Princeton University Press.

Steinkamp, M. W., & Maehr, M. L. (1984). Gender differences in motivational orientations toward achievement in school science: A quantitative synthesis. *American Educational Research Journal, 21*, 39–59.

Stiggins, R. J., & Conklin, N. F. (1992). *In teachers' hands.* Albany: State University of New York Press.

Supovitz, J. A., & Brennan, R. T. (1997). Mirror, mirror on the wall, which is the fairest test of all? An examination of the equitability of portfolio assessment relative to standardized tests. *Harvard Educational Review, 67*, 472–506.

Toenjes, L. A., Dworkin, A. G., Lorence, J., & Hill, A. N. (2000). *The lone star gamble: High-stakes testing accountability, and student achievement in Texas and Houston.* University of Houston, Department of Sociology, Sociology of Education Research Group.

U.S. Department of Education. (1991). *America 2000: An education strategy*. Washington, DC: Author.
Vinovskis, M. A. (1996). Analysis of the concept and uses of systemic educational reform. *American Educational Research Journal, 33*, 53–85.
Vinovskis, M. A. (1999). *The road to Charlottesville: The 1989 education summit*. Washington, DC: National Education Goals Panel.
Waltman, K. K., & Frisbie, D. A. (1994). Parents' understanding of their children's report card grades. *Applied Measurement in Education, 7*, 223–240.
Wells, A. S., & Crain, R. L. (1994). Perpetuation theory and the long-term effects of school desegregation. *Review of Educational Research, 64*, 531–555.
Wilder, G. Z., & Powell, K. (1989). *Sex differences in test performance: A survey of the literature* (Report No. 89-3). New York: College Entrance Examination Board.
Willingham, W. W., & Cole, N. S. (Eds.). (1997). *Gender and fair assessment*. Princeton, NJ: Educational Testing Service.