Working Paper #451
Princeton University
Industrial Relations Section
March 2001
http://www.irs.princeton.edu/pubs/working_papers.html

Would Smaller Classes Help Close the Black-White Achievement Gap?

Alan B. Krueger
Princeton University and NBER

and

Diane M. Whitmore
Princeton University

March 2001

This paper was prepared for a conference cosponsored by the Brookings Institute and Edison Schools, Inc. entitled, "Closing the Gap: Promising Approaches to Reducing the Achievement Gap." We thank Pat Turri and Jayne Zaharias for providing data, and David Card and Jens Ludwig for helpful discussions; they are not responsible for any mistakes we may have made.

Would Smaller Classes Help Close the Black-White Achievement Gap?

Executive Summary

This paper examines the effect of reducing class-size on student achievement, with particular attention to differential effects by race. A review of the literature suggests that low-income and black students tend to benefit more from attending a smaller class than white students. We extend the literature by providing new results from a long-term follow-up of students who participated in Tennessee's Project STAR. Project STAR was an experiment that randomly assigned 11,600 elementary school students and their teachers to a small class (target of 13-17 students), regular-size class (22-25 students) or regular-size class with a teacher-aide. The experiment began with the wave of students who entered kindergarten in 1985, and lasted for four years. After third grade, all students returned to regular-size classes. We analyze the effect of past attendance in a small class on standardized test scores through the eighth grade, on whether students took the ACT or SAT college entrance exam, performance on the ACT or SAT exam, criminal conviction rates, and teen birth rates.

The results indicate that, while students are in small classes, average test scores increase by 7-10 percentile points for black students and by 3-4 percentile points for white students. After all students are returned to regular-size classes in 4th grade, the gains from having attended a small class fall to about 5 points for black students and 1.5 points for white students, and persist at around that level. If all students were in a small class in grades K-3 for one to four years, we estimate that the black-white test-score gap would fall by 38 percent in grades K-3, and by 15 percent thereafter. Combining estimates of the effect of small classes on 3rd grade test scores from the STAR experiment with national trends in the pupil-teacher ratio for black and white students since 1971, we find that historical movements in the pupil-teacher ratio can account for almost all of the narrowing of the black-white test score gap as measured by the National Assessment of Educational Progress (NAEP) exam.

We also find that having attended a small class compared to regular-size class raises the likelihood that black students take the ACT or SAT college entrance exam from 31.8 to 41.3 percent, and raises the likelihood that white students take one of the exams from 44.7 to 46.4 percent. *As a consequence, if all students were assigned to a small class, the black-white gap in taking a college entrance exam would fall by an estimated 60 percent.* In addition, we find that past attendance in a small class raises the average score on the ACT or SAT exam by 0.15-0.20 standard deviation for black students, and by 0.04 standard deviation for white students.

Lastly, we find that the teen birth rate was one third less (3.2 versus 4.8 percent) for white females who were assigned to a small class than for those assigned to a regular-size class, and the fatherhood rate was 40 percent less (1.5 versus 2.5 percent) for black teenage males assigned to a small class than for those assigned to a regular-size class. The effect of class size on teenage births for other groups was not statistically significant. Black male students assigned to small classes were less likely to be convicted of a crime than those assigned to regular-size classes, but the effect was not statistically significant.

Introduction

The studies in Jencks and Phillips (1998) document the distressingly large gap in academic achievement between white and black students. On the 1999 National Assessment of Educational Progress (NAEP) exam, for example, the average 17-year-old black student scored at the 13th percentile of the distribution of white students on the math exam, and the 22nd percentile on the reading exam. Although part of this gap may result from racial biases in features of achievement tests, Jencks (1998) provides evidence suggesting that, at least in part, racial test score gaps reflect real differences in skills. Moreover, the black-white gap in academic achievement appears to be an important contributing factor to the black-white gap in income, health, crime, and other outcomes.[1] At the beginning of the last Century, W.E.B. Dubois correctly predicted that *the* problem of the 20th Century would be the color line. The problem of the 21st Century might well be the color line in academic achievement.

While the sources of the black-white achievement score gap are not well understood, it is clear that identifying strategies to help reduce the gap should be a priority for researchers and public policy makers alike. This paper considers the effect of reducing class size on student achievement. Contrary to the early work of the Coleman report and Hanushek's (1986, 1997) summaries of the literature, there is an emerging consensus that students' standardized achievement scores do increase as a result of attending smaller classes (see, e.g., Hedges, Laine and Greenwald, 1994 and Krueger, 2000). Resources matter. Nonetheless, lowering class size is one of the more costly educational interventions commonly considered. Our earlier cost-benefit analysis suggests that, when all students are considered, the benefits of smaller classes in terms of the monetary value of achievement gains are about equal to the additional costs, assuming an

[1] See, Neal and Johnson (1996), for example, for evidence on the impact of differential cognitive achievement on the black-white earnings gap.

initial class size level of 23 students and a discount rate of 5.5 percent (Krueger and Whitmore, 2001). A finding of a "normal rate of return" from reducing class size should not be a surprise or a disappointment because it suggests that local school are maximizing their allocation of resources. Moreover, reducing class size could still be a particularly desirable policy goal if disadvantaged groups benefit comparatively more from attending smaller classes.

Indeed, a small body of evidence, reviewed in Section 2, suggests that African American students tend to benefit more than white students from attending a smaller class. Project STAR, an experiment conducted in Tennessee in the late 1980s, is the only large-scale randomized experiment ever conducted to measure the effect of class size. Mosteller (1995) considered Project STAR "one of the most important educational investigations ever carried out and [it] illustrates the kind and magnitude of research needed in the field of education to strengthen schools." In Section 3 we present new estimates of the effect of reducing class size on standardized achievement scores, SAT and ACT test taking, crime and teenage pregnancy rates, separately for blacks and whites based on Project STAR. To place the magnitude of the gain from reducing classes by 7 students in context, in Section 4 we compare the effects of attending a smaller class to the gain found in recent voucher experiments for black students. Finally, in Section 5 we compare the observed reduction in the black-white test score gap over time as measured by the NAEP exam to the predicted narrowing of the gap based on the nationwide trend in the pupil-teacher ratio and the STAR estimates.

1. Trends in the Black-White Achievement Gap

To illustrate the dimensions of the problem, Figure 1 displays black and white 17-year-old students' average scores on the NAEP Math and Reading exam, normalized so the

nationwide score and standard deviation in 1996 both equal one.[2] Figure 2 displays the corresponding results for 9 year olds. Despite a small dip in the late 1980s, on this scale (which is admittedly quite wide) white students' achievement scores have been rather stagnant since the early 1970s. In the early 1970s, black students in both age groups scored about 1.2 standard deviations less on both the math and reading test than white students.

As Jencks and Phillips (1998) have emphasized, the NAEP data indicate that the black-white reading gap among 17 year olds has declined by almost half, and the math gap declined by almost a third. Nevertheless, a sizable gap remains – in 1999 black students scored 0.7 standard deviation lower on reading and 1.0 standard deviation lower on math. Also worrisome, the progress in narrowing the racial achievement gap appears to have stalled out since the late 1980s, and the gap has widened somewhat in the last five years.

Similar to the NAEP exam, data from the SAT exam show a narrowing of the black-white score gap since the 1970s, although the narrowing continued into the 1980s and the gap did not start to expand until the mid 1990s.[3] In 1976, for example, white students scored 1.16 standard deviations higher on the math portion and 1.08 standard deviations on the verbal portion. By 1991, the gap narrowed to 0.85 and 0.81 standard deviation. In 2000, using the re-centered exam, the gaps were slightly higher: 0.85 and 0.92 standard deviation. The SAT exam, of course, is not designed to be representative of the population, and is affected by changes in the composition of test takers, so the results are difficult to interpret. This problem aside, the SAT

[2] That is, we subtracted the 1996 national scale-score average from each year's score, and divided the resulting quantity by the 1996 cross sectional standard deviation. Differences between whites and blacks, and over time, can therefore be interpreted as changes relative to the 1996 standard deviation.

[3] SAT data by race for 1976-1995 are from the *Condition of Education – 1996*, Supplemental Table 22-2. Subsequent years are from various years of the College Board's *College-Bound Seniors National Profile Report*, available at www.collegeboard.org.

exam indicates that black students have made progress, although an enormous achievement score gap remains.

To gain some perspective on what it means for the distribution of black students' scores to be centered 0.9 standard deviation below that of other students, suppose a highly selective college only accepts students who score in the top five percent of the distribution, and that scores are normally distributed. Then black students would have to score 2.55 standard deviations above the mean of their distribution to be admitted – a hurdle cleared by only 0.5 percent of students. At an even more selective college, one that has a top one percent admissions cutoff, say, only 0.06 percent of black students would be admitted. If the mean test performance of black students increased by 0.2 standard deviations, the number of black students who clear this admissions hurdle would double. Raising the distribution of black students' test scores would greatly increase racial diversity at elite colleges.

2. Previous Literature

A common reaction to work on class size is, "Why bother to look at class-size effects? Hasn't Eric Hanushek definitively shown that hundreds of studies find no systematic relationship between class size and student achievement? Resources just don't matter." Krueger (2000) argues that this skepticism is unsupported by the research literature. First, Hanushek's latest tabulation of the literature is based on 59 articles on class size and 41 on expenditures per student, 22 of which were included in both; there are not hundreds of studies. Second, the literature seems larger than it actually is because Hanushek often extracts multiple estimates from the same paper, and then treats all estimates as separate, independent studies. Hanushek (1997), for example, extracted 277 estimates of the effect of class size from 59 different studies.

The number of estimates taken from each study varies widely: as many as 24 estimates were extracted from each of two papers (which used the same data set), and only one estimate was extracted from 17 studies apiece. Third, and most importantly, the number of estimates Hanushek extracted from a study is systematically related to the study's findings, with fewer estimates taken from studies that tend to find positive effects of smaller classes or greater expenditures per student.

Each estimate that Hanushek extracted was coded as positive, negative or unknown sign, and as either statistically significant or insignificant. The estimates were then tabulated (see column 1 of Table 1). A consideration of the way a few of the studies were treated illustrates some of the problems with this approach.

- Two studies by Link and Mulligan (1986 and 1991) each contributed 24 estimates, or 17% of all estimated class size effects. Both papers estimated separate models for math and reading scores by grade level (3rd, 4th, 5th, or 6th) and by race (black, white, or Hispanic), yielding 2 x 4 x 3 = 24 estimates apiece. One of these papers, Link and Mulligan (1986), addressed the merits of a longer school day using an 8% subsample of the data set used in their 1991 paper. Class size was only included in the regression specifications reported in the earlier paper as an interaction term with classmate ability levels, which generally pointed to a beneficial effect of attending a small class. Nevertheless, Hanushek coded 24 estimates as statistically insignificant, and unknown sign.

- Cohn and Millman (1975) estimated a series of education production functions using a sample size of 53 secondary schools in Pennsylvania. Hanushek selected 11 OLS estimates, 10 of which were negative, but excluded the authors' preferred 2SLS estimates, which corrected for simultaneity bias and were consistently more positive. In addition, the OLS estimates controlled for both the average class size in a high school and the pupil-teacher ratio, making the class-size variable difficult to interpret.

- Only one estimate was extracted from Summers and Wolfe (1977), who analyzed data for 627 sixth-grade students in 103 elementary schools. They mentioned that data were also analyzed for 533 eighth-grade students and 716 twelfth-grade students, with similar class-size results, but these results were not included in Hanushek's tabulation. Summers and Wolfe (1977; Table 1) provide two sets of regression estimates: one with pupil-specific school inputs and another with school-averages of school inputs. They also provide pupil-level estimates of

> class-size effects estimated separately for subsamples of low, middle, and high achieving students, based on students' initial test scores (see their Table 3). Yet Hanushek selected only one estimate from this paper – the main effect from the student-level regression. Why the estimates reported for the various subsamples were excluded is unclear.

Any reasonable standard would not place 11 times more weight on Cohn and Millman's study than on Summers and Wolfe's. By using estimates as the unit of observation, Hanushek implicitly weights studies by the number of estimates he extracted from them. There is no reason why study quality should be related to the number of estimates extracted. Indeed, some of the studies from which Hanushek extracted multiple estimates have estimated problematic specifications. For example, a dozen studies simultaneously controlled for expenditures per student and students per teacher. In such a specification, School A can only have a smaller class than School B by paying its teachers less or crimping on other resources – which is not the policy experiment most people have in mind when they think about reducing class size. [4]

Table 1 summarizes Krueger's (2000) reanalysis of Hanushek's literature review. The first column treats all estimates that Hanushek extracted equally. These results led Hanushek to conclude, "There is no strong or consistent relationship between school inputs and student performance."

When all *studies* are given equal weight, however, the literature does exhibit systematic evidence of a relationship between school inputs and student achievement. To weight the studies equally, in column (1) each study is assigned the proportion of estimates that are positive and significant, negative and significant, and so on, according to Hanushek's coding of the estimates, and then the arithmetic average is taken over all studies. Indeed, the number of studies that find positive effects of expenditures per student outnumbers those that find negative effects by almost

[4] These criticisms should not necessarily be interpreted as a critique of the underlying studies. Many of the studies were not about class size, and only conditioned on class size as an ancillary variable.

four to one. The number of studies that find a positive effect of smaller classes exceeds the number that find a negative effect by 57 percent. Differences of these magnitudes are unlikely to have occurred by chance.

We would argue that studies form a more natural unit of observation for this type of a literature summary than estimates because it is studies that are accepted for publication, not individual estimates. A weak paper can perhaps overcome the skepticism of a referee by including more estimates. In addition, most of the multiple estimates were either drawn from papers that used the same sample of students to examine different outcomes (math tests, reading tests, composite tests, etc.), so the estimates within a paper are not independent, or drawn from papers that carved up one sample into ever smaller subsamples, so the results were sta tistically imprecise. Finally, and probably most importantly, weighting all studies equally reduces the impact of researcher discretion in choosing which estimates to select.

To crudely (but objectively) assign more weight to higher quality studies, in column (3) the studies are assigned a weight equal to the 1998 "impact factor" of the journal that published the article, using data from the Institute for Scientific Information. The impact factors are based on the average number of citations to articles published in the journals in 1998. Impact factors are available for 44 of the 59 class-size studies in the sample; the other 15 studies were published in books, conference volumes, or unpublished monographs. Studies not published in journals were assigned the impact factor of the lowest ranked journal. The weighted mean of the percentages is presented in column 3 of Table 2. Although there are obvious problems with using journal impact factors as an index of study quality (e.g., norms and professional practices influence the number of citations), citation counts are a widely used indicator of quality, and the impact factor should be a more reliable measure of study quality than the number of estimates

Hanushek extracted.[5] The results are quite similar when either the arithmetic mean or journal-impact-weighted mean is used.

Columns 4-6 of Table 1 repeat this same exercise using Hanushek's tabulated results for expenditures per student. Here, studies that find a positive effect of school spending outnumber those that find a negative effect by nearly four to one. As a whole, we believe that when fairly summarized the literature does suggest that school resources matter – or at least it suggests that one should be less confident in the view that they don't matter.

Of particular interest is how class size affects achievement for black students. Table 2 summarizes the methods and findings of studies that provide separate estimates by race. The first five studies in Table 2 were included in Hanushek's survey. We also expanded the list by including some studies published after Hanushek's survey was written. We include a recent study by Stecher et. al. (2000), which evaluates the class-size reduction initiative in California; early work on the Project STAR class-size reduction experiment by Finn and Achilles (1990); Molnar, et al.'s (1999) study of Wisconsin's Sage program, and papers by Mora (1997) and Boozer and Rouse (2001), who use the National Educational Longitudinal Survey (NELS) data. One limitation of some of these studies for our purposes is that class size was often included as one variable in a "kitchen-sink" model because the authors did not intend for their estimated class size effect to receive attention. For example, Link and Mulligan (1991) controlled for both class size and hours of subject-specific instruction. If teachers in smaller classes can spend less class time on discipline and administration, one would expect that they spend more time instructing students on math and reading. Holding instruction time constant but varying class size confounds the estimated effect of smaller classes.

[5] Hanushek has argued that studies that use a "value added" specification are of the highest quality, but a number of authors have highlighted problems with the value added specification. See, for example, Ludwig and Bassi (1999),

To facilitate comparison, we scaled the coefficients to reflect the impact of a 7-student decrease in class size, which is the average class-size reduction in Project STAR. Where possible, we standardized the results to report test-score effects in terms of standard deviations, and reported standard errors. In some cases, however, the studies provided insufficient information to perform these calculations.

Although the findings vary considerably from study to study, on the whole the results suggest that attending a smaller class has a more beneficial effect for minority students than for non-minority students.

Stecher, et al. (2000) provides a particularly relevant evaluation because they considered the effects of an actual, state-wide class-size reduction initiative in California that could provide a model for other states. In this enormous education reform, which was championed by then-Governor Pete Wilson, California school districts that chose to participate received just over $800 for each K–3 student enrolled in a class of 20 or fewer students to encourage smaller classes. Because of the scale of this intervention, many implementation problems were encountered that do not arise in small-scale demonstration studies. For example, some higher income school districts reportedly raided teachers from lower income districts. In addition, many new classrooms had to be built to accommodate smaller classes, and temporary structures were often used. Nonetheless, Stecher, et al. find that, after two years, the California class-size reduction initiative led to a 0.10 S.D. increase in math scores and a 0.05 S.D. increase in reading scores on the SAT-9 exam for third graders. Both of these effect sizes were statistically significant. They also found that schools with a larger share of minority students had larger effect sizes. The effect size was 0.10 S.D. larger on the math exam in schools with 75 percent or

---

Cawley, Heckman and Vytlacil (1999) and Lindahl (2000).

more minority students compared to those with 25 percent or fewer minority students, for example, but this differential effect was not statistically significant.

Several other studies do not provide separate estimates by race, but do examine the effect of reduced class size on low-income or low-achieving students. For example, Hanushek, Kain and Rivkin (1998) find a positive effect of smaller classes on math and reading achievement of low-income 4^{th} and 5^{th} grade students in Texas (37 percent of the sample), but insignificant (mostly positive) effects for other students. Likewise, Grissmer, et al. (2000) report that the lowest-SES states with high pupil-teacher ratios (26 students per teacher) gain 0.17 standard deviation on test scores from a 3-student reduction in the pupil-teacher ratio. The effect size diminishes as average pupil-teacher ratio decreases, or as socio-economic status (SES) increases. And Summers and Wolfe (1977) find that low-achieving students perform worse when in larger classes, while high-achieving students perform better. Thus, the literature suggests that disadvantaged students – minorities, low-SES, and low-achievers – gain the most from smaller classes. This conclusion is also reinforced by the findings from the Project STAR experiment discussed below.

3. Results from Project STAR

The Tennessee STAR experiment is the best-designed experiment available to evaluate the impact of class size in the early grades on student achievement. In this section, we explore the impact on black and white students of attending a smaller class in the early grades. We are able to examine short-term achievement outcomes (e.g., test scores while the experiment was going on), long-term achievement outcomes (e.g., scores on the ACT or SAT exams nine years after the conclusion the experiment), and long-term non-academic outcomes, including teen births and incarceration rates.

Project STAR was an experiment in which a total of 11,600 students in kindergarten through 3rd grade were randomly assigned to a small class (target of 13-17 students), regular-size class (target of 22-25 students), or regular-size class with a full-time teacher's aide, within 79 Tennessee public schools.[6] The initial design called for students to remain in the same class type from grades K-3, although students were randomly re-assigned between regular and regular/aide classes in first grade. New students entering Project STAR schools in grades 1-3 while this cohort was participating in the experiment were randomly assigned to a class type. Students who left the school or repeated a grade were dropped from the sample being tracked during the experiment, although data on their subsequent performance in many cases was added back to the sample after 3rd grade, as other data sources were used. In 4th grade, all students were returned to regular classes.

Data are available for about 6,200 students per year in grades K-3, and about 7,700 students per year in grade 4-8, after the experiment ended. The average student in the experiment who was assigned to a small class in the experiment spent 2.3 years in a small class. An important feature of the experiment is that teachers were also randomly assigned to class types. Krueger (1999) evaluates some of the problems in the implementation and design of the STAR experiment, including high rates of attrition and possible nonrandom transitions between grade levels, and concludes that they did not materially alter the main results of the experiment.

It is important to emphasize that the small-class effects are measured by comparing students from different class-types in the same schools. Because students were randomly assigned to a class-type within schools, student characteristics – both measurable, such as free-

[6] See Word, Johnston, Bain, et. al (1990), Nye, Zaharias, Fulton, et al. (1994), Achilles (1999) or Krueger (1999) for more detail on the experiment.

lunch status, and unmeasurable, such as parental involvement in students' education – should be the same across class-types, on average.

A. *Standardized Test Scores, Grades K-8*

Because students were randomly assigned to small and regular classes within schools, it is important to control for school effects while estimating the "treatment effect" of being assigned to a small class. A simple estimator in this case is the "balanced-sample estimator," which holds the distribution of students across schools constant. Specifically, for each school and grade we calculated the average percentile rank for students assigned to small classes and those assigned to normal-size classes.[7] (Because earlier research found that students in regular classes with and without a teacher aide performed about equally well, we pool the aide and regular student together. We call this pooled group "normal-size" classes.) We then calculated the weighted average of these school-level means, using as weights in each case the number of normal-size-class students in the school that grade.[8] The difference between the two weighted means holds school effects constant. Intuitively, this arises because this difference could be calculated by first computing the difference in mean performance between small- and regular-class students *within each school*, and then taking the weighted mean of these school-level treatment effects.[9]

It is important to stress that class-type in this analysis is based on the class the student attended in his or her *initial* year in Project STAR, and does not vary over time. Using this

[7] The percentiles were derived in grades K-8 by using the distribution of raw scores for students in regular and regular/aide classes, as described in Krueger (1999). We use the average percentile score of the math and reading exams. If a student repeated a grade, we used his or her first test score for that grade level.

[8] For regular students, the resulting average, of course, is just the unweighted average score in the sample of regular students.

[9] Regressions of test scores on a dummy indicating initial assignment to a small class and school fixed effects yielded qualitatively similar results; see Krueger and Whitmore (2001).

categorization leads to what is commonly called an "intent to treat" estimator, as it was the intention to assign the students to their original class type. As a result, the estimated differences in means between small- and normal-class students are not subject to bias because of possible non-random transitions after the initial assignment. The estimates, however, will provide a lower bound estimate of the effect of *actually attending* a small class because some of the students assigned to a small class actually attended a normal-size class, and vice versa. This probably understates the effect of attending a small class by 10-15 percent.[10]

Figure 3a displays the weighted average percentile scores by class assignment for black students, and Figure 3b displays the corresponding information for white students. In each case, the weights were the number of regular size students *of that race* in the school. As Finn and Achilles (1990), Krueger (1999), and Krueger and Whitmore (2001) have found, the figure indicates that black students benefited comparatively more from being assigned to a small class. In grades K-3, black students in small classes outperformed those in normal-size classes by 7-10 percentile points, on average, while white students in small classes had a 3-4 percentile-point advantage over their counterparts in normal-size classes.[11] For both racial groups, the small-class advantage remained roughly constant during the course of the experiment – that is, through third grade. In terms of standard deviation units (of the full sample), black students gained about 0.26 of a standard deviation from being assigned to a small class while white students gained about 0.13 standard deviation.

In 4th grade, the effect size falls about in half for both racial groups, and remains roughly constant thereafter. The gain from having been assigned to a small class is approximately 5

[10] We derive this figure by regressing a dummy indicating whether students actually attend a small class on a dummy indicating whether they were initially assigned to a small class and school-by-wave dummies.

percentile points for black students in grades 4-8, and 1.5 points for white students in grades 4-8. These effects are still statistically significant at conventional significance levels. The decline in the treatment effects in 4th grade could result from several factors. As mentioned, after 3rd grade the experiment concluded and all students were enrolled in regular-size classes. Unfortunately, this also coincided with the time the assessment test was changed: in grades K-3, students took the Stanford Achievement Test, and in grades 4-8 they took the Comprehensive Test of Basic Skills (CTBS). Both tests are multiple-choice standardized tests that measure reading and math achievement, and are taken by students at the end of the school year. We suspect the change in the test is not critical for the reduction in the treatment effect because the year-to-year correlations in students' percentile rankings are about the same between 3rd and 4th grade (which encompass the different tests) and other adjacent grades (which use the same test). Another factor could be that the 4th grade sample is a subset of the overall sample because Memphis schools administered the CTBS test only to about one-third of their students in 1990; in later years the test was universally administered.

The composition of the students underlying Figures 3a and 3b is changing over time. The decline in the treatment effect in the first year after all students moved into regular-size classes is still apparent, however, if we just use the sub-sample of students with scores available in both 3rd and 4th grade. It is also apparent if we just use the subset with scores available in both 3rd and 5th grade, which avoids possible problems created by the omission of many Memphis 4th graders.

Although it is tempting to interpret the decline in the treatment effect between 3rd and 4th grade as a fading out of the gains achieved by the small-class students, it is also possible that

[11] The estimated intent-to-treat effects are unlikely to have occurred by chance. In grades K-3, the standard error of the small-regular difference is around 1.2 for black students and 0.9 for white students. In grades 4-8 the standard errors fall to around 1.2 for black students and 0.8 for white students.

peer effects increased the performance of students who had been in normal-size classes relative to those who had been in small classes after the experiment ended.

Another issue to bear in mind in interpreting the trends over time in Figures 3a and 3b is that these tests are scaled according to percentile ranks, and that percentile ranks are not a cardinal measure. It is possible – even likely – that a given percentile gap corresponds to a larger educational difference in later grades than in earlier grades. [12] The discrete decline in the small-class advantage the year all students moved into regular size classes suggests, however, that something real happened. Nevertheless, a bottom line result from Figures 3a and 3b is that assignment to a small class led to greater improvement in the relative position in the distribution of test scores for minority students than for white students – and it did so during the years when students were in small classes and subsequently.

The following calculation suggests that the effect of assignment to a small class on the racial test-score gap is sizable. In 3rd grade, for example, the black-white gap in the average percentile rank was 18.8 points in normal-size classes and 11.7 points in small classes. So according to these figures assigning *all* students to a class of 15 students as opposed to 22 students for a couple of years in grammar school would lower the black-white gap by about 38 percent. Part of this beneficial effect appears to fade out over time – by 8th grade, assignment to a small class in the early grades appears to narrow the black-white test score gap by 15 percent. These figures are likely to understate the benefit of attending a small class because, as mentioned previously, not everyone assigned to a small class attended one.

Lastly, we can also calculate the average treatment effect for white students using the school-level weights for black students to infer whether white students who attend the same mix

[12] Finn et al. (1999) present evidence that – when grade-equivalent scores are used to scale the tests – the gap between students in small and regular-size classes expands from grades K-3, and from grades 4-8.

of schools as black students have an average treatment effect that is close to that found for black students. Interestingly, for grades K-3, that is what the data suggest – the average treatment effect for white students, weighted by the number of black students in the school, is close to what Figure 3a shows for black students. After grade three, however, this exercise produces an even smaller estimate for white students than what we found in Figure 3b, where the treatment effects for whites are weighted by the number of white students. If we do this exercise for black students (i.e., weight their school-level treatment effects by the number of white students in normal classes in the school), the impact of being assigned to a small class is uniformly smaller in all grades.

B. College-Entrance Exam Taking Rates

a. Updated ACT data

Students who were assigned to a small class in grades K-3 were significantly more likely to take the SAT or ACT college-entrance exam. Krueger and Whitmore (2001) reported initial results using data for students who graduated high school in 1998. That paper found that small-class attendance raised the likelihood that black students take the ACT or SAT by a quarter – from 31.7 to 40.2 percent. As a result, the black-white gap in the test-taking rate was 54 percent smaller in small classes than in regular classes. This work, however, was limited by incomplete data on test scores. In the remainder of this section, we present more complete results encompassing several additional years of ACT data, and focus on racial differences.

To create the original longitudinal database with SAT and ACT information used by Krueger and Whitmore (2001), in the summer of 1998 the ACT and ETS organizations matched Project STAR student data to their national database of test records. The match was performed

using student names, dates of birth, and Social Security numbers, or two of the three identifiers if one field was missing. [13] Because the test files are organized by graduating class, at that time it was only possible to match test data for students who graduated in the class of 1998. As a result, any Project STAR student who repeated a grade (or skipped ahead) and did not graduate on schedule with the class of 1998 could not be matched. Based on data through 8th grade, it appears that nearly 20 percent of the Project STAR sample had been left behind a grade. Any student who had been left behind would be categorized as not having taken the ACT or SAT, even if the student had already taken the test during their junior year or took it a year later. The resulting misclassification tends to attenuate the effect of small classes on test-taking rates. To minimize the effects of misclassification, our earlier work presented most results limiting the sample to those who graduated on schedule.

For this paper we have obtained additional ACT data to augment the sample to include those students who did not graduate on schedule. The records that were not matched by ACT in our earlier data set were re-submitted, and ACT attempted to match them to their databases for the classes of 1997 - 2000. The SAT match was not updated, but this is probably inconsequential because based on the first match only 3.5 percent of Project STAR test takers took the SAT and not the ACT. Further, students who took only the SAT tended to have stronger academic records, so they are less likely to be behind a grade. As before, records were matched on student name, date of birth and Social Security number. The match was done nationwide when all three variables were present. For cases that lacked a valid Social Security number (32 percent), the match was restricted to the state files of Tennessee, Kentucky and Mississippi. These three states accounted for 95 percent of the ACT matches in the first round.

[13] See Krueger and Whitmore (2001) for a more complete description. After the records were merged, student names, dates of birth and Social Security numbers were concealed to preserve confidentiality.

In the new match, an additional 10.7 percent of previously unmatched students were linked to ACT data.[14] Several checks indicate that the data were linked properly for students who were matched. For example, the correlation between the students' ACT score percentile rank and their 8th grade CTBS percentile rank was 0.74, which is similar to the correlation between other percentile scores of tests given four years apart.[15] In addition, the sex of the student based on their Project STAR record matched their ACT-reported sex in 97.5 percent of the matches. These checks suggest that the Project STAR data were linked correctly, and show that the new match is about the same quality as the previous match.

C. Test Taking Rates

To examine whether assignment to a small class influences the college-entrance exam test-taking rate, we again use a balanced sample estimator to adjust for school effects.[16] For each school denoted j, racial group denoted r, and class type (small, regular, regular with aide) denoted c we estimate the percent of students who took either the ACT or SAT exam, denoted $\overline{Y}_{jr}^{C}$. We then calculated the weighted average of the school means, using as weights the number of regular-class students in each school (N^{R}):

[14] Many of the students appear to have been held back a grade: 48.7 percent of students matched reported that they graduated high school in 1999 or 2000, one or two years late. Another 10.8 percent graduated in 1997 – a year ahead of normal progress. The class of 1998 – which should have been matched in the original data – accounted for 38.9 percent of the new match. Of the new class of 1998 matches, 43 percent took the test after graduating high school, while the remaining 57 percent appear to have been missed in the first round, in part reflecting better information on Social Security numbers that we obtained in the meantime.

[15] The correlation between the 3rd grade Stanford Achievement Test and 7th grade CTBS is .75, and the correlation between the CTBS in 4th and 8th grade is .80.

[16] We note that nominally, beginning in the Spring of 1998, Tennessee required high school students to take an exit exam as part of state-wide curriculum changes introduced by the Education Improvement Act. Students completing the university-track diploma were required to take the SAT or ACT. Students opting for a technical diploma could take the SAT, ACT or Work Keys. Despite this new requirement, however, the share of Tennessee high school students taking the ACT did not increase in 1998, according to ACT records. Moreover, students who were not college bound would be likely to take the Work Keys test, which we do not code as taking a college entrance exam. Thus, we suspect our measure provides a meaningful indication of whether students were college bound, despite this requirement.

$$\overline{\overline{Y}}_r^C = \frac{\sum_j \overline{Y}_{jr}^C N_{jr}^R}{\sum_j N_{jr}^R}.$$

Notice that the difference between the small-class and regular-class means $(\overline{\overline{Y}}_r^S - \overline{\overline{Y}}_r^R)$ can be written as $\sum_j (\overline{Y}_{jr}^S - \overline{Y}_{jr}^R) \bullet w_{jr}$ where w_{jr} is the fraction of all students in the experiment of that race who were assigned to regular-size classes in that school. This shows that the difference between the means can be written as the weighted average of the school-level treatment effects, which makes transparent that school effects are held constant.

Our findings are illustrated in Figure 4. This figure reports the percent of students by race who took either the SAT or ACT, by the type of class they attended during their first year in Project STAR. For white students, Figure 4 indicates that 46.4 percent of students initially attending small classes took a college-entrance exam, compared to 44.7 percent in regular classes and 45.3 percent in regular/aide classes. These differences in rates are not statistically significant. Black students were substantially more likely to take the SAT or ACT if they were assigned to a small rather than regular-size class: 41.3 percent of black students assigned to small classes took at least one of the college entrance exams, compared with 31.8 percent in regular classes and 35.7 percent in regular/aide classes. The chance of such a large difference in test-taking rates between the small and regular class students occurring by chance is less than one in 10,000.

To interpret the magnitude of these effects, note that the black-white gap in taking a college entrance exam was 12.9 percentage points for students in regular-size classes, and 5.1 percentage points for students in small classes. Thus, assigning all students to a small class is estimated to reduce the black-white gap in the test-taking rate by an impressive 60 percent.

As before, we can also calculate the average treatment effect using different weights to infer whether white students who attend a similar mix of schools as black students have the same treatment effect as black students. That is, we have calculated the weighted average of the school-level treatment effects for whites using the number of black students in the school as weights. Remarkably, this indicated an even larger treatment effect for white students than was previously found for black students – about an 11 percentage-point higher rate of test taking for small-class students than normal-size-class students. Likewise, if we use the school-level treatment effects for the black students and the number of white regular students in the school as weights, the treatment effect for blacks shrinks to about that found for whites in Figure 4.

These findings suggest that small classes matter for blacks because of something having to do with the schools they attend, rather than something inherent to individual black students per se. For example, it is possible that black students attend schools that have a disproportionately high number of disruptive students, or students with special needs, which distracts their teachers from instructional time.[17] In this case, white students in those schools would also benefit from smaller classes.

D. ACT Test Scores, With and Without Selection Adjustment

Next, we examined the scores the students attained on the ACT and SAT exams. For students who took the SAT but not the ACT exam, we converted their SAT score to an ACT-equivalent score using a concordance developed jointly by ACT and the College Board.[18] For

[17] See Lazear (1999) for a formal economic model that predicts that smaller classes lead to higher achievement by reducing the number of disruptions in a class.

[18] See www.collegeboard.org for the concordance. The concordance maps re-centered SAT I scores (verbal plus math) into ACT composite scores. 121 students – or 2.6 percent of the test-taking sample – took the SAT and not the ACT. For the 378 students in our sample who took both tests, the correlation between their SAT and ACT scores is 0.89.

any student who wrote the ACT exam we used the ACT score even if he or she also took the SAT. For students who took an exam more than once we used the first score. Naturally, any analysis of ACT and SAT scores can only be performed on the subset of students who took one of the exams. This creates a potential selection problem. Because a higher proportion of students from small classes took the SAT or ACT exam, it is likely that the group from small classes contains a higher fraction of relatively weak students. That is, stronger students are likely to take an exam regardless of their class assignment, but marginal students who are induced to take the exam because they attended a small class are likely to be lower-scoring students. Such a selection process would bias downward the effect of attending a small class on average test scores. The bias is also likely to be greater for black students, because a higher share of black students were induced to take the exam as a result of attending a small class.

To simplify the analysis, we compare students who initially attended small classes to the combined sample of those who initially attended either regular or regular/aide classes, and we control for school effects instead of school-by-entry-wave effects. Also, because we later implement a Heckman (1976) selection correction, we use raw ACT scores instead of percentile ranks for this analysis. The raw ACT scores in our sample range from 9 to 36 and are approximately normally distributed.

The results are reported in Table 3. For the sample of test takers, the average ACT score was virtually identical for students who were assigned to small and normal-size classes. The average white student in a small class scored 19.88, while the average white student in a regular class scored 19.87. Black students in small classes averaged 16.3, while black students in regular classes scored 16.1. The differences between small and normal-size classes are not statistically significant.

Past studies of state-level data have shown that average test scores tend to decline when more students take a college entrance exam, most likely because the marginal test takers are weaker students than the average student (see, e.g., Card and Payne, 1998). In Project STAR, there were two confounding effects: selection and treatment. One might expect the treatment to result in small-class students scoring slightly higher on the ACT, as they did on previous tests through the 8th grade. But since a larger percentage of students assigned to small classes took the exam, a larger share of weaker students in small classes likely took the test. As a result, it is difficult to interpret the score results because scores are only reported conditio nal on taking the exam, and the treatment appears to have affected the likelihood of taking the exam – particularly for black students. Columns (2) and (3), and (5) and (6) present two types of estimation results that attempt to adjust for this sample selection problem.

Columns (2) and (4) present results of a standard Heckman-correction procedure for white and black students. Identification in this model is based solely on the assumption of normal errors, as there is no exclusion restriction. We also calculate the "effect size" by dividing the coefficient on the small-class dummy by the standard deviation of ACT scores among all students who took the exam (equal to 5.4). The Heckman correction doubles the point estimate on the effect of attending a small class for white students, but the coefficient is still statistically insignificant and qualitatively small. For blacks, however, column (5) indicates that after adjusting for selection, students in small classes score 0.15 standard deviation higher th an those in regular classes.

In columns (3) and (6) we present results from a different approach for adjusting for selection. Here we have artificially truncated the sample of students from small classes so that the same proportion of students from small and regular-size classes is represented in the test-

taking sample. Specifically, we drop from the sample the bottom X percent of students based on their test results, where X is determined so that the share of students from small classes who took the exam equals the share from regular-size classes. This approach is valid if all the additional small-class students induced to take the ACT are from the bottom of the distribution, and if attending a small class did not change the ranking of students in small classes. Although the former assumption is extreme, the results should provide an upper bound on the impact of selection bias, and are an interesting point of comparison to the Heckman-correction results.

In Krueger and Whitmore (2001) we provided some diagnostic information on these two selection-correction approaches by comparing the Heckman-correction procedure and the linear truncation model for eighth-grade students, where we had test scores for the full universe of students. If we artificially truncated the sample to those who later took the ACT or SAT exam, we found that the two selection correction procedures bracketed the effect estimated by OLS for the full sample.[19]

The results in columns (3) and (6) are quite similar to the Heckman-correction results in columns (2) and (5). For white students, the linear truncation and Heckman-selection-correction procedure indicate that students in small classes score insignificantly differently from students in normal size classes, with a point estimate corresponding to a 0.04 standard deviation. For black students, the linear-truncation procedure yields and effect size of 0.20 standard deviations, somewhat larger than the 0.15 effect size from the Heckman-correction procedure.

[19] In principle, the Heckman procedure provides an estimate of the effect of attending a small class on test scores for the entire population of students (including those who did not take the test), whereas the linear-truncation approach provides an estimate of the effect of attending a small class on scores for students from regular classes who otherwise would have taken the ACT. If there is a homogeneous treatment effect, the two parameters would be equal.

E. The Effect of Class Size on Other Outcomes

By raising economic and educational opportunities, smaller classes may also indirectly affect the frequency of negative social outcomes such as crime, welfare receipt and teen pregnancy. Here we present initial results on the effect of smaller classes on criminal convictions and the teen birth rate.

The criminal conviction data come from Tennessee State Department of Corrections records, and were matched to Project STAR using student Social Security numbers. Crimes for which an individual was not convicted are not counted in the data set. Also, because the match was only performed in Tennessee, any crime committed by a student in another state is not included in the data set. This measurement problem would lead to a downward-biased estimate of the difference in criminal behavior by class assignment if the same proportion of participants from small and large classes moved out of state, and if those students are just a random sample of the small and large class students. As long as small-class assignment did not increase the probability that a family moved away from Tennessee, the measurement error will likely attenuate the small-class impact.

Criminal convictions are in this sample are rare: only 1.6 percent of Project STAR students overall were reported as being convicted of a crime, and 2.6 percent of males were. Since 88 percent of those convicted were males, for this analysis we restricted the sample to include only males. We employed the balanced-sample estimator described above, and report the results in Table 4. In the first row, we measure the rate of criminal activity of males by assigning a one to any student who was matched to the crime conviction data, and a zero otherwise. Columns (3) and (6) display the balanced-sample estimator. In column (6), black

males in small classes are 0.6 percentage point less likely to be convicted of a crime than those in normal-size classes. This difference, however, is not close to being statistically significant.

Sentence length is measured as the maximum sentence (in days) faced by individuals for their specific crimes. Data are not available on length of actual sentence or time served, but maximum sentence length provides a measure of the severity of the crime committed. The sentences range from one year for minor theft and drug offenses to 8-12 years for aggravated robbery and serious drug offenses. Students without convictions were assigned a zero sentence length. Column (6) indicates that black males in small classes on average committed crimes that carried 12 fewer days (or 24 percent) of maximum prison time than their peers in larger classes back in elementary school. Despite the fact that this effect is sizable, it is not statistically significant with this size sample. Class size has a much smaller and opposite-signed effect on both crime rates and sentence length for white males.

Another important outcome to measure is the teen birth rate. Births to teens are highly correlated with female high school dropout rates and welfare utilization. Maynard (1997) reports that roughly four-fifths of teen mothers end up on welfare, and their children are more likely to have low birth weight. Hotz, McElroy and Sanders (1999) estimate that giving birth as a teen reduces the probability that a girl will graduate from high school by 15 to 16 percent. The bottom portion of Table 4 presents results on the effect of small-class assignment on the teen birth rate.

Birth records, like crime records, were matched in the State of Tennessee only. Records were matched by Social Security number of the mother and father reported on the birth certificate to Project STAR records, and then matches were confirmed by comparing student

name.[20] If both of a newborn child's parents were STAR students, the birth record is counted for both the mother and the father. Birth records were only available by calendar year. We restricted our analysis to births during 1997 and 1998 because most students graduated high school in 1998.

The birth rates were constructed as follows: aggregate births counts by class type, race, gender and school were provided from Tennessee records. These were converted to rates by dividing by the total population in each cell. Row (3) in Table 4 reports birth rates for females by race and class-assignment type. As shown in column (3), small class assignment is associated with a statistically significant 1.6 percentage point (or 33 percent) lower teen birth rate for white females. Row (4) reports similar results for births in which a male STAR student was reported to be the father according to the birth records; the lower fatherhood rate for black males from small classes is on the margin of statistical significance. The effect of class-assignment on the teen birth rate for white males and black females is not statistically significant.

4. Comparison to Voucher Results

It is helpful to put the Project STAR class-size results for African Americans in context by comparing them to other interventions. Here we compare the effect of attending a smaller class to the effect of private school vouchers, as estimated by Howell, Wolf, Peterson and Campbell (2000). They report short-term test-score gains estimated from privately funded voucher experiments conducted on low-income students in grades 2-8 in Dayton, Ohio, New York City, and Washington, D.C. These experimental results indicated that after their first year in a private school, test scores increase by an average of 3.3 percentile points, on average. By

[20] Individual data were then aggregated by initial school, class type, race and gender. We do not have access to the micro-data.

the end of the students' second year, black students who switched to a private school scored an average of 6.0 percentile points higher than their counterparts who did not switch, when the three sites are aggregated by weighting the site-effects by their inverse sampling variances. For other ethnic groups, test scores declined on average if students switched to a private school, although the decline was not statistically significant. Howell, et al. also find that the test-score effect for blacks does not vary by subject matter; math and reading improved about the same amount.

To compare these voucher results to class-size reduction, we estimated a regression model for all black students who were in their second year after having been assigned to a small class, regardless of their grade level. Thus, the sample consists of students in their second year of Project STAR, whether that was grade 1, 2 or 3. For this sample, we estimated an OLS regression in which the dependent variable was the average Stanford Achievement Test percentile score. The explanatory variables included a dummy variable for *initial assignment* to a small class, school-by-entry-wave dummies, current-grade dummies, free-lunch status, and sex. The effect of having been assigned to a small class in this model is 7.9 points (with a standard error equal to 1.1).

To further improve the comparability of the samples, we estimated the same model described in the preceding paragraph for black students who were initially on *free or reduced price lunch*, as the voucher experiment was restricted primarily to those on free lunch. For this sample, assignment to a small class raised scores by an estimated 8.6 percentile points (standard error =1.2) after two years.[21]

As mentioned previously, our use of initial assignment to a small class understates the effect of attending a small class by about 15 percent because not everyone assigned to a small class actually attended one. Likewise, not every student randomly provided a voucher switched

to a private school, but the estimates from Howell et al. reported here adjust for incomplete take-up. Their "intent-to-treat" estimate in the second year after assignment is 3.5 percentile points.

We conclude from this comparison that, when comparable samples are considered, black students who attended a small class for two years in the STAR experiment improved their test performance by around 50 percent more than the gain experienced by black students who attended a private school as a result of receiving a voucher in the New York, Dayton and Washington voucher experiments.[22]

5. Class Size and the Reduction in the Black-White Gap over Time

Historically, black students attended schools with far larger pupil-teacher ratios than did white students. Horace Mann Bond (1934) eloquently summed up the situation this way: "Negro schools are financed from the fragments which fall from the budget made up for white children." Throughout the 20th Century, the gap in school resources between white and black school has narrowed (see, e.g., Boozer, Krueger and Wolkon, 1992). In 1915, for example, the pupil-teacher ratio was 61 in black schools in the segregated states and 38 in white schools; by 1953-54, on the eve of *Brown vs. Board of Education*, it was 31.6 in black schools and 27.6 in white schools. By 1989, Boozer, Krueger and Wolkin estimate that the pupil-teacher ratio had converged in the schools the average black and white student attended, although Boozer and Rouse (2001) provide evidence that black students still attended larger classes within the same schools in the early 1990s.

As mentioned previously, the black-white test score gap has also narrowed over the last 30 years – and this trend probably began long before 1970, although consistent, nationally

---

[21] Math and reading are impacted by the same magnitude if the equation is estimated separately.

[22] We note that it is also possible that some of the gain for students who attended a private school may result from the fact that, at least in New York City, participating private schools had 2 to 3 fewer students per class than public schools, on average.

representative data are not available.[23] The pupil-teacher ratio fell over this period for both black and white students, and it fell slightly more for black students. Can the decline in the racial test score gap recorded in the NAEP be explained by the contemporaneous reduction in average class sizes, especially among black students?

To calculate the effect of reduced class sizes on the test-score gap, we obtained the national average pupil-teacher ratio by year from the Digest of Education Statistics. Although average class size and the pupil-teacher ratio are not the same quantities, they should be closely related. Blacks were, on average, in larger classes than whites during most of the 1970s, but the gap closed by the late 1980s. To capture the relative difference in class-size reduction over this period, we crudely adjusted the national pupil-teacher ratio to derive separate estimates by race. For each year, we assigned the national pupil-teacher ratio to whites, and inflated the pupil-teacher ratio by the inverse of the relative white/black pupil-teacher ratio reported by Boozer, Krueger and Wolkon (1992) to obtain an estimate for blacks.

To estimate the effect of a one student reduction in class size, we used student-level data from STAR to estimate a Two Stage Least Squares (2SLS) model by race. The dependent variable in the second-stage equation a student's average scale score on the 3rd grade math and reading sections of the Stanford Achievement Test, and the key endogenous regressor was the student's actual 3rd grade class size. We also controlled for whether the student was assigned to a class with a teacher aide, gender, free-lunch status, and school dummies. We instrumented for class size using a dummy variable indicating whether the student was initially assigned to a small or normal size class (as well as the other exogenous regressors). We then scaled the estimated

[23] In a careful decomposition of the narrowing of the black-white NAEP test score gap between 1970 and 1988, Cook and Evans (2000) find that at least three quarters of the reduction in the gap occurred within schools and no more than one quarter occurred because of relative changes in parental education. The fact that black students

effect of a one student change in class size by the standard deviation of test scores for all regular-size students in 3rd grade. This yielded an estimate that a one student reduction in class size would lead to an increase in the 3rd grade test score by 0.02 standard deviation for whites and by 0.05 standard deviation for blacks. We use these estimates, together with the change in pupil-teacher ratios over time, to predict the gap in test scores by race.

Figures 5a and 5b display the actual (based on NAEP) and predicted black-white test score gap in math and reading, scaled by the subject-specific standard deviation in 1996. First consider Figure 5a, which shows the decline in the math test score gap among 9-year-olds between 1973 and 1999. The actual score gap declined by 0.21 standard deviation, from 1.04 to 0.83 standard deviation, over these 26 years. During this period, average class sizes in elementary schools fell from 23.0 to 18.6 for whites, and from 25.4 to 18.6 for blacks. Based on our calculations, these changes in class sizes are predicted to reduce the black-white test score gap by 0.25 standard deviation, slightly more than the observed decline. Overall, however, the correspondence between the actual and predicted decline in the black-white gap is remarkably close. Figure 5b shows the analogous results for reading tests of 9-year-olds. Again, the predicted narrowing in the black-white achievement score gap is closely mirrored by our prediction based on changes in class size for black and white students over this period. It is also interesting to note that most of the predicted narrowing in the test score gap came about because of the decline in the pupil-teacher ratio generally – which has a larger effect on black students according to our estimates – than from the larger decline in the pupil-teacher ratio for black students relative to white students.

---

appear to benefit more from smaller classes than white students – combined with the general decline in class size – could account for the large within-school effect that they find.

We would not push these results too far, however, because other important determinants of student achievement scores have also changed since the 1970s, and because we can only crudely measure average class size for white and black students. Nevertheless, the results do suggest that the effect of class-size on achievement by race as estimated from the STAR experiment are roughly consistent with the trends in test scores and pupil-teacher ratios that we have observed in the aggregate over time.

6. Conclusions

To summarize, our analysis of the STAR experiment indicates that students who attend smaller classes in the early grades tend to have higher test scores while they are enrolled in those grades than their counterparts who attend larger classes. The improvement in relative ranking on standardized tests that students obtain from having attended a small class is reduced when they move into regular-size classes, but an edge still remains. Moreover, black students tend to advance further up the distribution of test scores from attending a small class than do white students, both while they are in a small class and afterwards. For black students, we also find that being assigned to a small class for an average of two years in grade K-3 is associated with an increased probability of subsequently taking the ACT or SAT college entrance exam, and 0.15-0.20 standard deviation higher average score on the exam. These findings are more or less consistent with most of the available literature.

Because black students' test scores appear to increase more from attending a small class than do white students', the decline in the pupil-teacher ratio nationwide over the last century should have led to a reduction in the black-white achievement gap. Moreover, the fact that the pupil-teacher ratio declined relatively more for black students should provide an added boost to

the reduction in the achievement gap. Our calculations suggest that the decline in the pupil-teacher ratio for black and white students experienced in the last 30 years can account for most of the reduction in the black-white achievement score gap, although other factors surely were at work as well.

An important question is, Why? Why do black students appear to gain more from attending a smaller class than white students? Although there is a need to look further inside this box, our analysis suggests that something about the schools black students attend leads to a greater impact of smaller classes. That is, white students who attend the same mix of schools as black students appear to profit from smaller classes by about as much as black students do, and *vice versa* for black students who attend predominantly white schools. More generally, we find that students who attended schools with lower average test scores in the elementary grades benefit the most from attending smaller classes. One possible explanation for these findings is that teachers have to move very slowly through the curriculum if they have weak students – e.g., because they are disrupted frequently or have to explain the material multiple times to the slower students – but if they have smaller classes they can effectively teach more material. By contrast, teachers in schools with well-behaved, self-motivated students can move quickly through the material regardless of class size. This type of an explanation might also partially explain why some countries, such as Japan, have high test scores despite having large classes. Regardless of the explanation, our findings suggest that class size reductions will have the biggest bang for the buck if they are targeted to schools with relatively many minority students. But if such targeting is politically infeasible, then reducing class size generally would still lead to a reduction in the black-white test score gap.

References

Achilles, C. (1999). *Let's Put Kids First, Finally: Getting Class Size Right*. Thousand Oaks, CA: Corwin Press.

Bond, Horace Mann, *The Education of the Negro in the American Social Order*. New York: Prentice Hall, 1934.

Boozer, Michael, Alan Krueger and Shari Wolkon. 1992. "Race and School Quality since *Brown* v. *Board of Education*." *Brookings Papers on Economic Activity: Microeconomics*, Martin N. Baily and Clifford Winston, eds., pp. 269-326.

Boozer, Michael and Cecilia Rouse. 2001. "Intraschool Variation in Class Size: Patterns and Implications." Forthcoming in *Journal of Urban Economics*.

Card, David, and Alan B. Krueger. 1992. "School Quality and Black-White Relative Earnings: A Direct Assessment." *Quarterly Journal of Economics* 107(1): 151-200.

Card, David and Abigail Payne. 1998. "School Finance Reform, the Distribution of School Spending and the Distribution of SAT Scores." U.C. Berkeley, Center for Labor Economics, Working Paper, forthcoming, *Journal of Public Economics*.

Cawley, John, James Heckman and Edward Vytlacil. 1999. "On Policies to Reward the Value Added by Educators." *Review of Economics and Statistics* 81(4), 720-27.

Cohn, E. and S. D. Millman. 1975. *Input-output Analysis in Public Education*. Cambridge, MA: Ballinger.

Cook, Michael and William Evans. 2000. "Families or Schools? Explaining the Convergence in White and Black Academic Performance," *Journal of Labor Economics* 18(4), October, pp. 729-54.

Ehrenberg, R. G. and D. J. Brewer. 1994. "Do School and Teacher Characteristics Matter? Evidence from 'High School and Beyond." *Economics of Education Review*, 13(1), pp. 1-17.

Finn, Jeremy D., and Charles M. Achilles. 1990. "Answers and Questions About Class Size: A Statewide Experiment." *American Educational Research Journal* 27 (Fall): 557-77.

Finn, Jeremy D., Susan Gerber, Charles M. Achilles and Jayne Boyd-Zaharias. 1999. "Short- and Long-term Effects of Small Classes." SUNY Buffalo, mimeo.

Grissmer, David W., Anne E. Flanagan, Jennifer Kawata and Stephanie Williamson. 2000. "Improving Student Achievement: What State NAEP Test Scores Tell Us." RAND Issue Paper 924, July.

Hanushek, Eric A. 1986. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature* 24 (September): 1141-77.

Hanushek, Eric A. 1997. "Assessing the Effects of School Resources on Student Performance: An Update." *Educational Evaluation and Policy Analysis* 19(2): 141-64.

Hanushek, Eric A., John F. Kain and Steven G. Rivkin. 1998. "Teachers, Schools, and Academic Achievement." NBER Working Paper 6691, August 1998.

Heckman, James. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement*, vol. 5, pp. 475-92.

Hedges, Larry V., Richard Laine, and Rob Greenwald. 1994. "Does Money Matter? A Meta- Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes." *Education Researcher* 23(3): 5-14.

Hotz, V. Joseph, Susan Williams McElroy and Seth G. Sanders. 1999. "Teenage Childbearing and Its Life Cycle Consequences: Exploiting a Natural Experiment." NBER Working Paper 7397, October.

Howell, William G., Patrick J. Wolf, Paul E. Peterson and David E. Campbell. 2000. "Test-Score Effects of School Vouchers in Dayton, Ohio, New York City and Washington, D.C.: Evidence from Randomized Field Trials." Program on Education Policy and Governance Research Paper, August.

Jencks, Christopher. 1998. "Racial Bias in Testing." In Jencks and Phillips, eds., *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press, pp. 55-85.

Jencks, Christopher and Meredith Phillips. 1998. *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press.

Krueger, Alan B. 1999. "Experimental Estimates of Educational Production Functions." *Quarterly Journal of Economics* 114(2): 497-532.

Krueger, Alan B. 2000. "Economic Considerations and Class Size." Princeton University Industrial Relations Section Working Paper 477, www.irs.princeton.edu, September 2000.

Krueger, Alan B., and Diane M. Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence From Project STAR." *Economic Journal* 111: 1-28.

Lazear, Edward P. 1999. "Educational Production." Working Paper No. 7349. Cambridge, Mass.: National Bureau of Economic Research.

Lindahl, Mikael. 2000. "Home versus School Learning: A New Approach to Estimating the Effect of Class Size on Achievement." Swedish Institute for Social Research, Vol. 43, *Studies of Causal Effects in Empirical Labor Economics*, University of Stockholm.

Link, Charles R., and James G. Mulligan. 1986. "The Merits of a Longer School Day." *Economics of Education Review* 5(4): 373-81.

Link, Charles R., and James G. Mulligan. 1991. "Classmates' Effects on Black Student Achievement in Public School Classrooms." *Economics of Education Review* 10(4): 297-310.

Ludwig, Jens and Laurie Bassi. 1999. "The Puzzling Case of School Resources and Student Achievement." *Educational Evaluation and Policy Analysis*, 21(4): pp. 385-403.

Maynard, Rebecca. 1997. *Kids Having Kids*. Washington, DC: Urban Institute Press.

Molnar, Alex, et al. 1999. "Evaluating the SAGE Program: A Pilot Program in Targeted Pupil-Teacher Reduction in Wisconsin." *Educational Evaluation and Policy Analysis* 21:2 (Summer): 165-177.

Mosteller, Frederick. 1995. "The Tennessee Study of Class Size in the Early School Grades." *The Future of Children: Critical Issues for Children and Youths* 5 (Summer/Fall): 113-27

Mora, Marie T. 1997. "Attendance, Schooling Quality, and the Demand for Education of Mexican Americans, African Americans, and Non-Hispanic Whites." *Economics of Education Review* 16(4): 407-418.

Neal, Derek, and William Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differentials." *Journal of Political Economy* 104 (October): 869-95.

Nye, Barbara, Jayne Zaharias, B.D. Fulton, et al. 1994. "The Lasting Benefits Study: A Continuing Analysis of the Effect of Small Class Size in Kindergarten Through Third Grade on Student Achievement Test Scores in Subsequent Grade Levels." Seventh grade technical report. Nashville: Center of Excellence for Research in Basic Skills, Tennessee State University.

Sengupta, J. K. and R. E. Sfeir. 1986. "Production Frontier Estimates of Scale in Public Schools in California." *Economics of Education Review*, 5(3), 297-307.

Stecher, B. M. and G. W. Bohrnstedt. 2000. *Class size reduction in California: The 1998-99 Evaluation Findings*. Sacramento, CA: California Department of Education, August.

Summers, Anita and Barbara Wolfe. 1977. "Do Schools Make A Difference?" *American Economic Review*, 67 (4), pp. 649-52.

Winkler, D. 1975. "Educational Achievement and School Peer Group Composition." *Journal of Human Resources*, 10(2), 189-204.

Word, E., J. Johnston, Helen Bain, et. al. 1990. "The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Technical Report 1985-1990." Tennessee State Department of Education.

Table 1: Reanalysis of Hanushek's (1997) Literature Summary; Studies of Class Size and Expenditures per Pupil

	Class Size			Expenditures per Student		
Result	**Weighted by No. of Estimates Extracted (1)**	**Equally-Weighted Studies (2)**	**Studies Weighted by Journal Impact Factor (3)**	**Weighted by No. of Estimates Extracted (4)**	**Equally-Weighted Studies (5)**	**Studies Weighted by Journal Impact Factor (6)**
Positive & Stat. Sig.	14.8%	25.5%	34.5%	27.0%	38.0%	40.1%
Positive & Stat. Insig.	26.7%	27.1%	21.2%	34.3%	32.2%	28.0%
Negative & Stat. Sig.	13.4%	10.3%	6.9%	6.7%	6.4%	6.3%
Negative & Stat. Insig.	25.3%	23.1%	25.4%	19.0%	12.7%	8.3%
Unknown Sign & Stat. Insig.	19.9%	14.0%	12.0%	12.9%	10.7%	17.3%
Ratio Positive to Negative	1.07	1.57	1.72	2.39	3.68	4.66
P-Value	0.500	0.059	0.034	0.0138	0.0002	0.0001

Notes: Columns (1) and (4) are from Hanushek (1997; Table 3), and implicitly weight studies by the number of estimates that were taken from each study. Columns (2), (3), (5) and (6) are from Krueger (2000). Columns (2) and (5) assign each study the fraction of estimates corresponding to the result based on Hanushek's coding, and calculate the arithmetic average. Columns (3) and (6) calculate a weighted average of the data in column (2), using the journal impact factors as weights. A positive result means that a small class size or greater expenditures are associated with improved student performance. Columns (1) - (3) are based on 59 studies, and columns (4) - (6) are based on 41 studies. P-value corresponds to the proportion of times the observed ratio, or a higher ratio, of positive to negative results would be obtained in 59 or 41 independent Bernouli trials in which positive and negative results were equally likely.

Table 2: Summary of Class-Size Studies that Provide Separate Estimates by Race

Study	Description	Findings for a 7-student reduction in class size
Ehrenberg & Brewer (1994)	Uses High School and Beyond individual-level data to look at effects of teachers and school resources on gain in test scores and on dropout behavior. Both equations include pupil-teacher ratio, expenditure per student, base-year test score, student gender, family income and size, parents' education level, a dummy variable indicating whether the school is in an urban setting, the percentage of students in the school that are black, the percentage that are hispanic, the percentage that are low-income, the difference between the percentage of black (hispanic) faculty and black (hispanic) students, and teachers' experience, education level, and quality of their undergraduate institution. The dropout equation is estimated by a probit model, and the gain model is estimated by OLS using a Heckman correction for dropouts. Median sample size is 1003.	Change in test score gain (standard errors in parenthesis): Blacks: 0.140 (0.108) Whites: 0.038 (0.029) Change in dropout rate (standard errors in parenthesis): Blacks: - 0.322 (0.358) Whites: - 0.007 (0.023)
Link and Mulligan (1991)	Estimates separate OLS regression models for individual-level CTBS math and reading scores for 3rd, 4th, 5th and 6th grade using the Sustaining Effects data set. Explanatory variables include pre-test score, class size, gender, hours of instruction, a dummy variable indicating whether the teacher recommends compensatory education, same race percentage of classmates, racial busing percentage, and the mean and standard deviation of classmates' pre-test scores. Median sample size is 6023.	Average change in test score, scaled by standard deviation (standard errors in parenthesis): Blacks: - 0.001 (0.014) Whites: - 0.004 (0.007)
Winkler (1975)	Examines the effect of the racial composition of a school on 8th grade Stanford reading scores using student data from a large urban California school district in 1964-65. The model also includes aggregate student/teacher ratio in grades 1-8, student IQ measurec in 1st grade, number of siblings, number of cultural items in home, parents' homeownership status, teacher salary, total administrative spending, share of peers with low-SES status, the change in low-SES status between elementary and middle school, the share of black students and the change in share between elementary and middle school, and share of teachers from prestigious colleges. Median sample size is 387.	Change in test score for 8 years of class-size reduction, scaled by standard deviation (standard errors in parenthesis): Blacks: 0.117 (0.156) Whites: 0.166 (0.170)
Card and Krueger (1992)	Estimates the effect of pupil-teacher ratio on returns to education. Uses Census data on wages in 1960, 1970 and 1980 for Southern-born men aggregated to state-by-cohort cells, linked to school characteristics in segregated states from 1915 to 1966. Uses weighted least squares to estimate by race the effect on return to education of pupil-teacher ratio, and state, year and cohort dummy variables. Sample size is 180.	Change in payoff to 1 year of education (standard errors in parenthesis): Blacks: 0.410 (0.347) Whites: 0.219 (0.389)
Sengupta and Sfeir (1986)	Sample contains 50 school-level observations on 6th graders in California. Dependent variables are math, reading, writing and spelling test scores. Explanatory variables are average teacher salary, average class size, percent minority, and interaction between percent minority and class size. Half of the 8 models also control for non-teaching expenditures per pupil. Estimates translog production functions by LAD.	Change in average test score, scaled by standard deviation: Blacks: 0.711 Whites: - 0.411

Table 2: Summary of Class-Size Studies that Provide Separate Estimates by Race

Study	Description	Findings for a 7-student reduction in class size
Stecher et al. (2000)	Evaluates the statewide class-size reduction in California enacted in 1996-97, which aimed to reduce the average class size in grades K-3 from 28 to no more than 20. Uses 3rd grade Stanford Achievement test data for reading, math, language and spelling. Presents differences between 3rd grade scores in schools with and without class-size reduction, after controlling for underlying differences in schools' (untreated) 5th grade scores.	Overall effect size is 0.073 standard deviation. Math and language tests have somewhat larger effect sizes in schools with high percentages of minority, low-income, and English-learner students.
Mora (1997)	Uses individual-level NELS data and runs logit models to examine the effect of school quality on propensity to drop out. Students were in 8th grade in 1998, and 1990 follow-up data were used to measure dropout status. Explanatory variables include pupil/teacher ratio, salary expenditures per pupil, length of school year, enrollment, school programs (counseling, departmentalized instruction, and GPA-requirement for activities), dummy for private school, location categories, student SES characteristics, and classroom racial and SES composition. Adjusts pupil-teacher ratio to measure average daily attendance per teacher, not total enrollment per teacher. Median sample size is 6677.	Change in probability of dropping out (standard error in parenthesis): Blacks: 0.012 (0.028) Whites: -0.067 (0.011)
Finn and Achilles (1990)	Reports results from a statewide experiment in Tennessee, in which students were randomly assigned to small or regular-size classes. Individual-level data on test scores in first grade are analyzed. School fixed-effects are also employed as explanatory variables. Median sample size is 3300.	Change in test score, scaled by race-specific standard deviation: Blacks: 0.254 Whites: 0.123
Boozer and Rouse (2001)	Find that class size often varies within school due to compensatory class assignments that put gifted and regular students in larger classes than special needs and remedial students. With compensatory resource allocation, they find that using school-level pupil teacher ratios may bias coefficient estimates downward. Individual-level NELS data are used to estimate the effect of class size on test score gains.	Overall, 7-student decrease in class size increases test scores by 0.49 standard deviation. Class size does not statistically significantly vary by race.
Molnar, et al. (1999)	Evaluates Wisconsin's SAGE program, which reduced pupil-teacher ratio in selected schools. Only schools with student poverty rates of 30 percent or more were eligible to apply. Uses 1st grade CTBS test data for reading, math and language arts. Presents differences between two cohorts of 1st grade students' scores in SAGE schools and comparison schools with similar characteristics. Sample size is 3944.	Change in test score, scaled by comparison group standard deviation: Blacks: 0.361 Whites: -0.127

Notes: Logit coefficients in Mora (1997) are transformed assuming the mean dropout probability equals 0.1.

Table 3: Effect of Class Size on ACT or SAT Score with and without Selection Correction
Dependent variable equals ACT or ACT-equivalent score

	White Students			Black Students		
Explanatory Variable	**No correction**	**Heckman correction**	**Linear truncation**	**No correction**	**Heckman correction**	**Linear truncation**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Intercept	20.233	16.386	20.242	17.073	7.443	17.164
	(0.138)	(0.524)	(0.138)	(0.275)	(3.610)	(0.274)
Small Class	0.009	0.209	0.206	0.213	0.834	1.079
	(0.169)	(0.210)	(0.167)	(0.204)	(0.266)	(0.203)
Female (1=yes)	0.056	1.787	0.021	0.522	2.229	0.378
	(0.156)	(0.197)	(0.156)	(0.190)	(0.237)	(0.191)
Free Lunch (1=yes)	-1.434	-4.859	-1.385	-1.715	-3.529	-1.725
	(0.180)	(0.241)	(0.179)	(0.265)	(0.332)	(0.263)
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	3198	7124	3173	1427	4117	1357
Effect Size	0.002	0.039	0.038	0.039	0.153	0.198

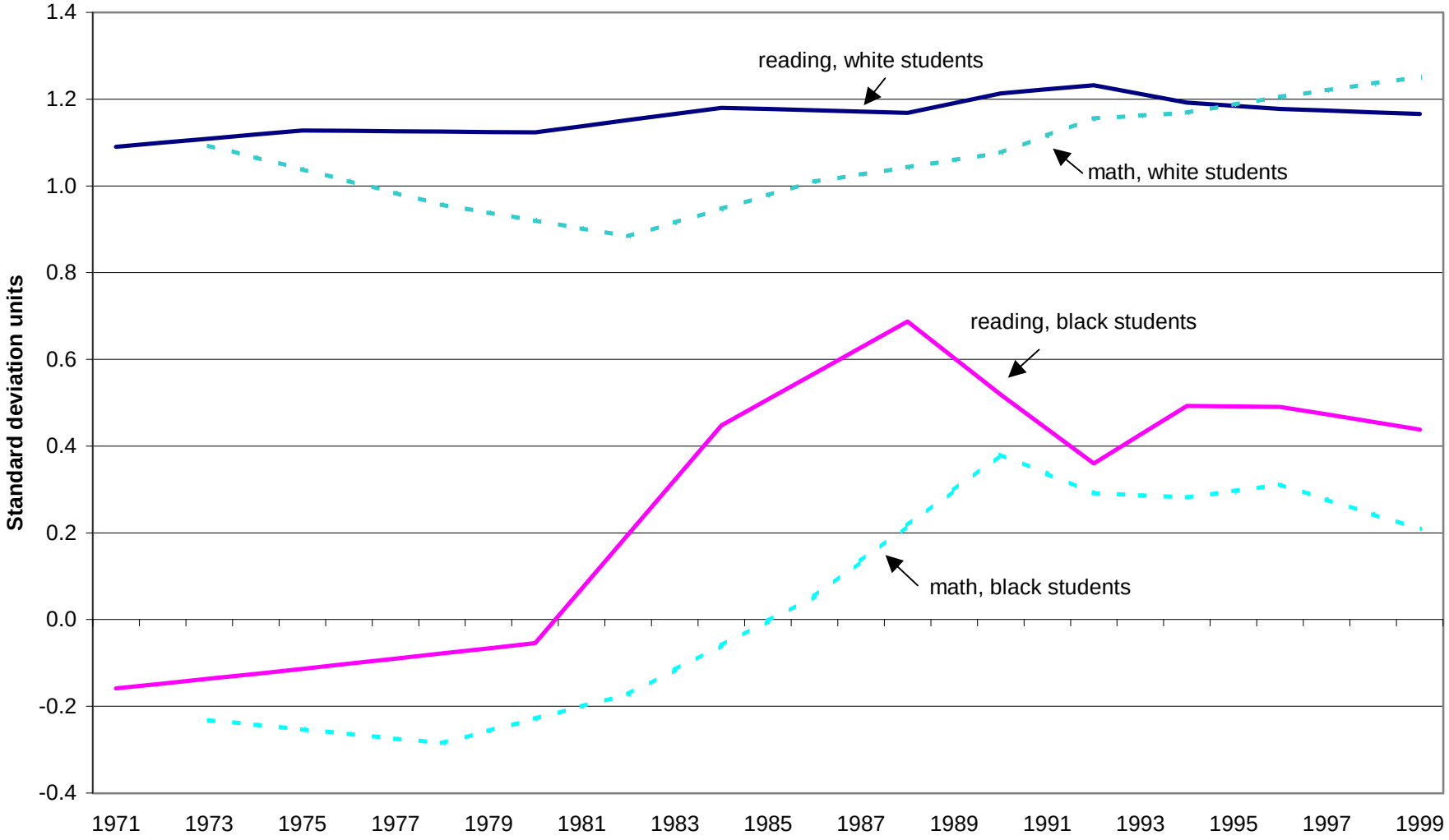
Note: Heteroskedasticity-adjusted standard errors are reported in parentheses for columns (1), (3), (4) and (6). If a student took only the SAT, that score is converted to its comparable ACT score (see text for details). The mean (standard deviation) of the dependent variable in column (1) is 19.9 (4.5), 19.9 (4.4) ir column (3), 16.1 (3.5) in column (4), and 16.3 (3.5) in column (6). The effect size is the coefficient on small divided by the standard deviation of test scores among the full sample of students (5.4).

Table 4: Effects of Small Classes on Crime and Teen Pregnancy

Dependent variable:	White			Black		
	Small Class	Normal Class	Difference	Small Class	Normal Class	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
(1) Ever convicted of a crime (males only)	0.023	0.022	0.001	0.025	0.031	-0.006
	(0.005)	(0.003)	(0.005)	(0.029)	(0.008)	(0.030)
(2) Average sentence length in days (males only)	26.7	24.4	2.3	37.7	49.9	-12.2
	(5.4)	(3.6)	(6.5)	(7.4)	(11.8)	(13.9)
(3) Birth rate (females only)	0.032	0.048	-0.016	0.059	0.044	0.015
	(0.006)	(0.004)	(0.007)	(0.010)	(0.005)	(0.011)
(4) Fatherhood rate (males only)	0.020	0.016	0.004	0.015	0.025	-0.010
	(0.002)	(0.004)	(0.005)	(0.004)	(0.005)	(0.006)

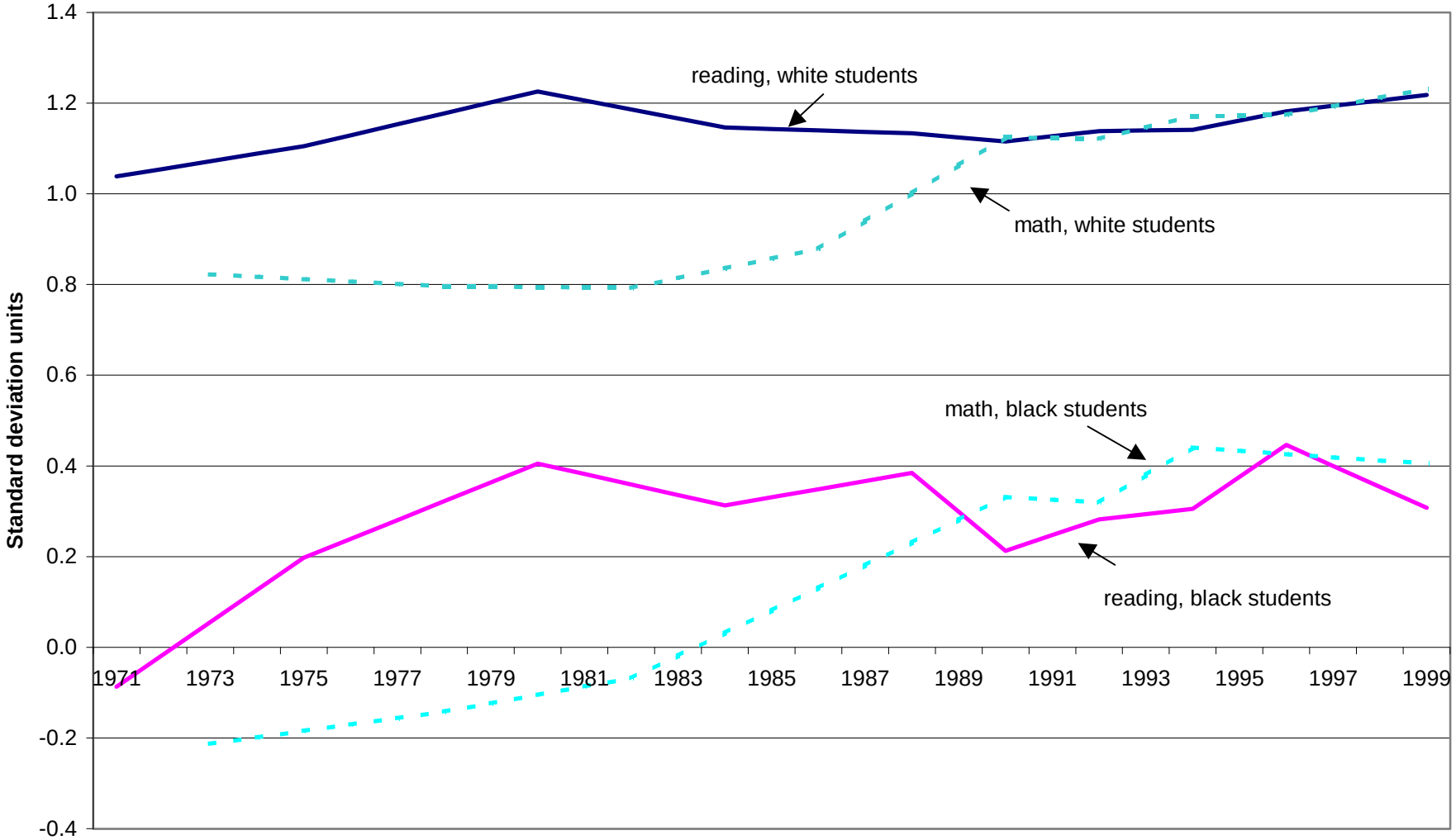
Note: Standard errors in parentheses. Balanced-within-school estimator used. Birth rates limited to births in 1997 and 1998.

Figure 1: Trends in Average Reading and Math Scores by Race, 17 Year Olds



Note: Data are from the National Center of Education Statistics. Scores are expressed as the raw score less the 1996 subject-specific mean raw score for all 17 year-olds, then this difference is divided by the 1996 cross-sectional standard deviation for all 17 year-olds, and 1.0 is added to the resulting normalized score.

Figure 2: Trends in Average Reading and Math Scores by Race, 9-Year Olds



Note: Data are from the National Center of Education Statistics. Scores are expressed as the raw score less the 1996 subject-specific mean raw score for all 9 year-olds, then this difference is divided by the 1996 cross-sectional standard deviation for all 9 year-olds, and 1.0 is added to the resulting normalized scores.

Figure 3a: Black Students' Average Test Scores using the Balanced Sample Estimator

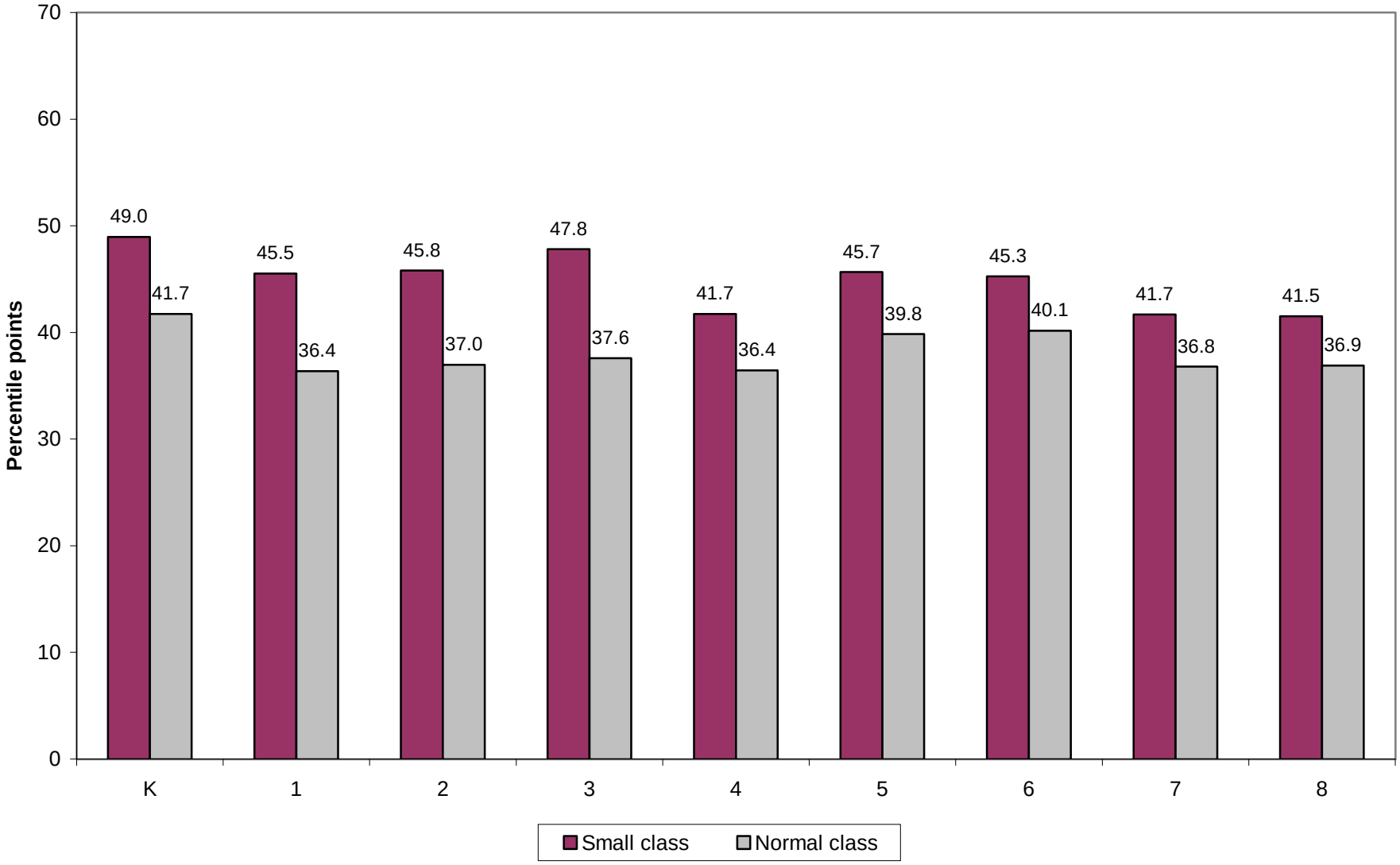


Figure 3b: White Students' Average Test Scores using the Balanced Sample Estimator

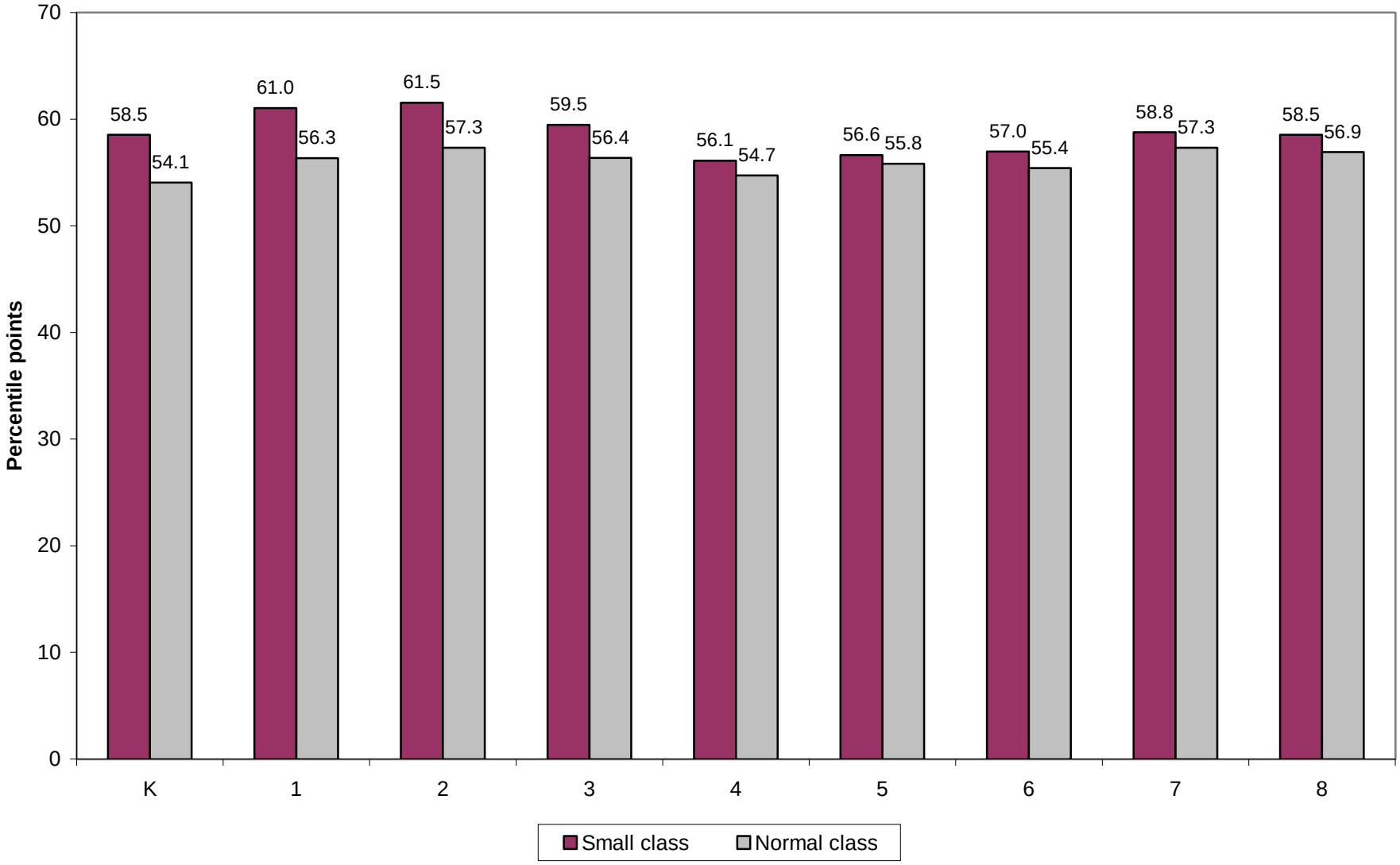
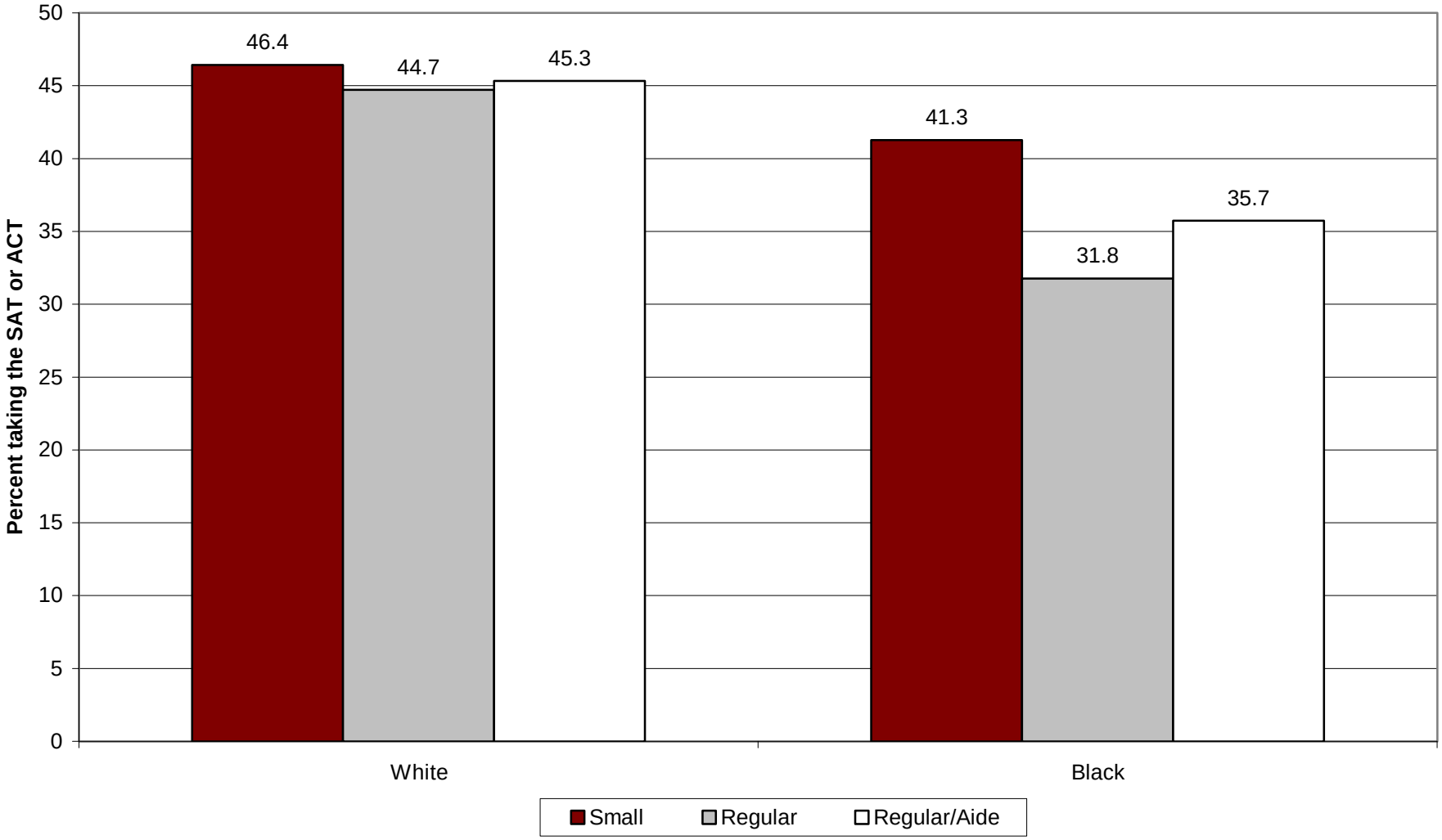
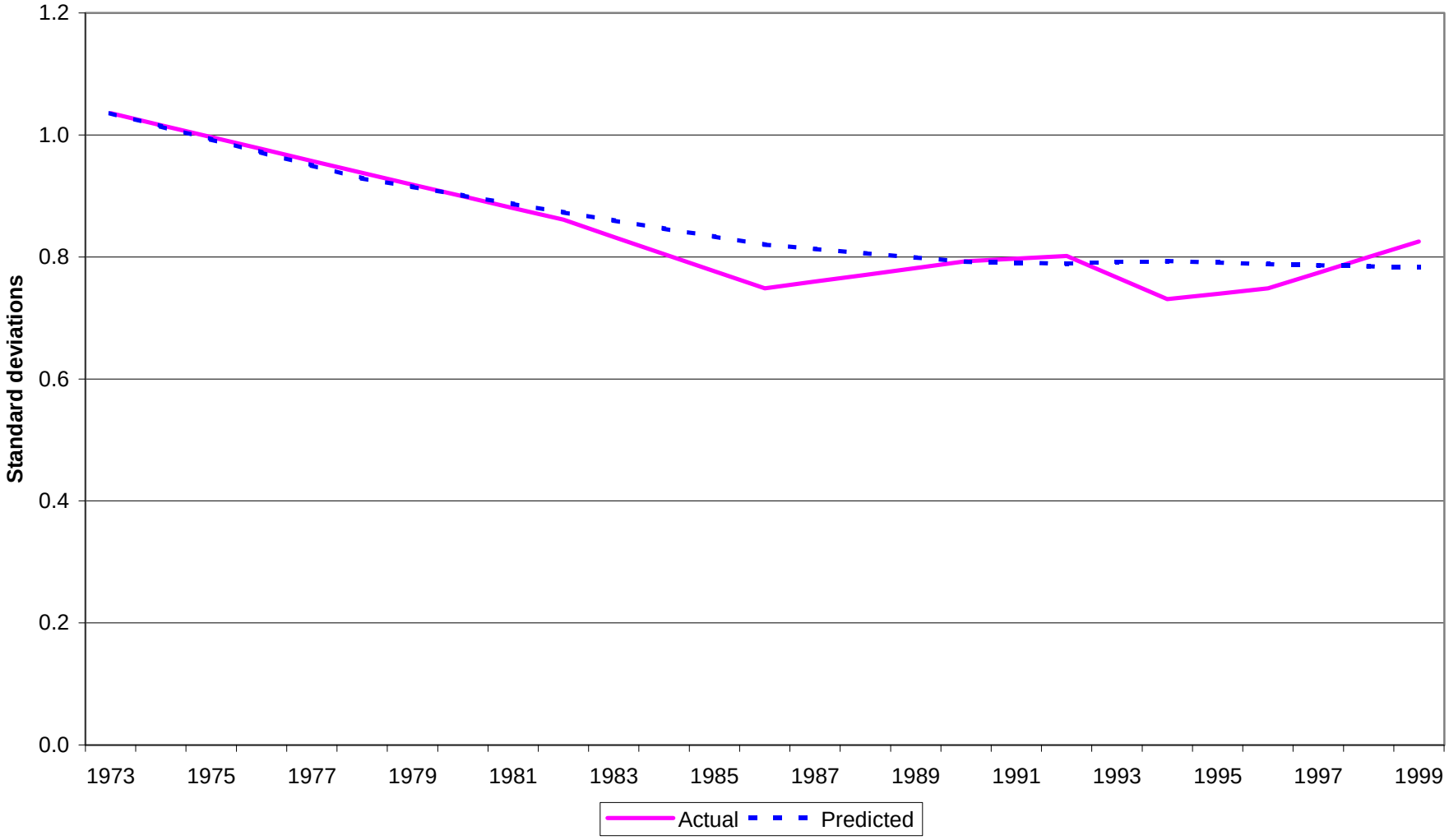


Figure 4: Percent of Students Taking the SAT or ACT by Initial Class Type



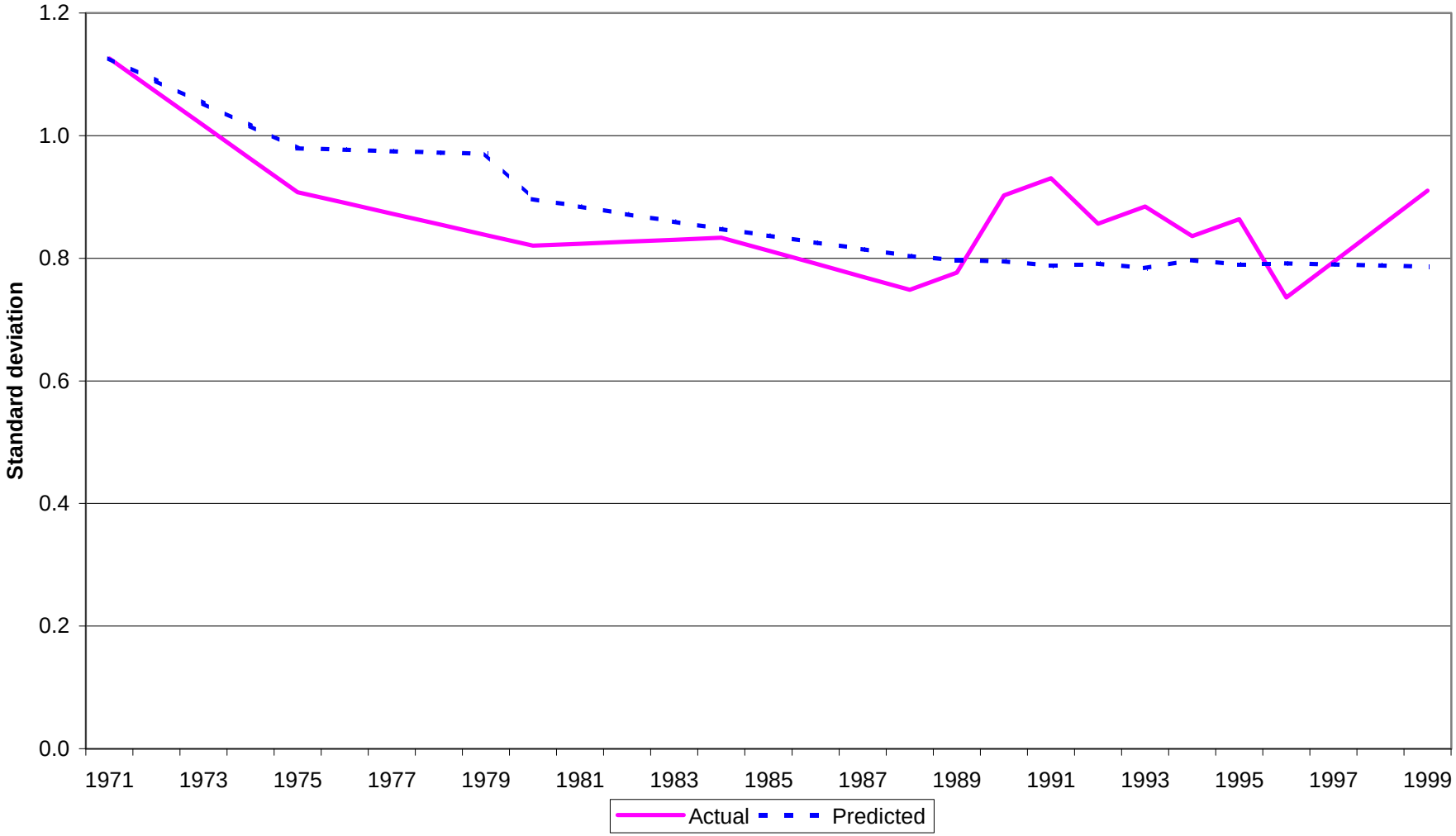
Note: Means are balanced within school using the balanced sample estimator described in the text.

Figure 5a: Black/White Gap in 4th-Grade Math Test Scores



Note: Predicted gap is estimated using the change in average elementary-school pupil-teacher ratio and black/white difference in pupil-teacher ratio from the Digest of Education Statistics and Boozer, Krueger and Wolkon (1992), and estimates the effect of reduced class size on test scores based on Project STAR data as described in the text.

Figure 5b: Black/White Gap in 4th-Grade Reading Test Scores



Note: Predicted gap is estimated using the change in average elementary-school pupil-teacher ratio and black/white difference in pupil-teacher ratio from the Digest of Education Statistics and Boozer, Krueger and Wolkon (1992), and estimates the effect of reduced class size on test scores based on Project STAR data as described in the text.