Racial Segregation and the Black-White Test Score Gap: Data Appendix

David Card and Jesse Rothstein

January 2007

This Appendix accompanies "Racial Segregation and the Black-White Test Score Gap," dated January 2007 and resubmitted to the Journal of Public Economics. We refer occasionally to the working paper version, NBER Working Paper #12078, dated March 2006.

This appendix describes the data sources used in our analysis, the processing of the original data into the MSA-level data set used for our main analyses, and the construction of our dependent and independent variables.

I. Data Sources

Our data on SAT scores come from samples of SAT-takers from the 1998-2001 cohorts of high-school graduates. Our data include 100% of black test takers from these cohorts. For white test takers, we observe 100% of those from California and Texas and 25% of all others. Test takers of other races (including "Hispanic," which is treated as a separate race on the SAT questionnaire) and those who did not report a race are not used. For each SAT taker, we observe the test score and a code for the high school attended. We also observe responses to the Student Descriptive Questionnaire, administered as part of the SAT. This questionnaire is the source for our variables characterizing parental education and family income for SAT takers.

For information about the size and racial composition of public schools, we use data from the Common Core of Data (CCD) schools survey. This is an annual census of public schools; we use the 1986/87 through 2000/01 waves. Each wave contains information on the school's enrollment at each grade level and on the racial composition of the school's overall enrollment.[1] We also use the CCD for our measures of teachers per pupil and per pupil expenditures (in Table 8 of the working paper) and free lunch rates in Table 7.[2] Our information about private schools comes from the 1997-98 Private School Survey (PSS).

Most of our information about MSA demographic characteristics comes from the 2000 Census. We use these data in a variety of forms. First, we use MSA-level tabulations from the Summary File 3. These are our source for metropolitan-level population, racial composition, average household income, and inequality of household income. Second, we use MSA-by-race tabulations from the Summary File 4, extracting records for white non-Hispanics (who did not list any other race) and for black non-Hispanics (who may have listed other races). From these, we compute black-white gaps in family structure and family income. Third, we use tract-level tabulations from the

[1] In recent years, the CCD has begun to provide information about grade-level racial compositions, but we do not use this. Racial composition data are unavailable in some states in some years. We discuss below our strategy for dealing with this, which involves the substitution of available data from alternate years.

[2] Free lunch rates are missing in many states. These are excluded from our analysis of this variable.

Summary File 1 to measure neighborhood segregation. Fourth, we use data from the public use microdata samples (PUMS). We define two universes in the PUMS data. First, we extract all children 0-17 living in metropolitan areas and merge them to their parents to obtain information about maternal and paternal education and wages. Second, we measure school enrollment and attainment for 16-24 year olds.

II. Data Processing

A. Merging

All of our analyses are conducted at the MSA level. We use consistent MSA definitions for all data sets, drawn from the U.S. Office of Management and Budget's 1999 classification.[3] We treat each PMSA within a larger CMSA as a distinct metropolitan area.

The 2000 Census Summary Files contain MSA codes. The PUMS files also contain MSA codes, but these are missing for some respondents. The smallest geographic unit reported in the PUMS is the Public Use Microdata Area (PUMA), which need not lie wholly within an MSA; when a PUMA spans an MSA boundary, the MSA code is set to missing. We assign these PUMAs to MSAs when a single MSA contains a majority of the PUMA's population. Some MSAs cannot be identified in this way (for example, when a majority of every PUMA appearing in the MSA is non-metropolitan). We eliminate all such MSAs from our data. When we compute averages for children aged 0-17 and their parents, we assign the families to the MSA where they lived in 2000. Our analyses of 16-19-year-olds, however, assign them to their location as of 1995.

Assigning schools in the CCD and PSS to MSAs is less straightforward. Although the CCD Agency file—with observations on school districts—contains an MSA code, this is frequently incorrect. Outside of New England, MSAs consist of whole counties, so we can use county codes (on the CCD Agency file and the PSS) to assign MSAs. For New England, we used mapping software and Census TIGER/Line files to overlay zip codes and MSA boundaries, assigning each zip code to the MSA or non-metropolitan area containing the majority of its land area. We used the resulting zip code-MSA crosswalk to assign New England schools to MSAs. As zip codes change over time and as schools sometimes have zip codes that contain no housing units, not all zip codes appeared in the crosswalk. When the school zip code could not be assigned, we next tried the agency zip code (for public schools). When this too failed, we coded schools by hand, using the district name and the address of both the school and district to assign each school to a town and thereby to an MSA. The crosswalks for each part of this operation are available from the authors.

The final step in the merging was to assign SAT takers to schools. The SAT data contain a school code, but the coding system differs from that used by the CCD and PSS. We obtained a file from the Educational Testing Service with the SAT school codes, school names and addresses, and (for some schools) CCD/PSS school codes. Using the names and addresses, we verified the crosswalk for those schools that already had CCD/PSS codes and extended it, where possible, to schools that were missing codes.

[3] This is available at http://www.census.gov/population/estimates/metro-city/99mfips.txt.

Although the resulting file is not complete, it successfully assigns upwards of 97% of public-school SAT-takers to schools, and approximately 70% of non-public-school SAT-takers. Both rates are higher in SAT states (discussed below), in the 80-90% range for private schools depending on the state. SAT-takers that cannot be assigned are discarded.

B. Measurement of Segregation

We measure segregation at both the tract and the school level. For tract-level segregation, we compute the fraction black or Hispanic in each tract in each MSA in 2000. We then compute the MSA-level average of this, weighting first by the number of blacks in the tract and then by the number of whites. The difference between these is our residential segregation measure.

Our school-level segregation measure attempts to capture the lifetime exposure of our cohorts of SAT-takers. To eliminate noise in the school-level enrollment counts, we average over several years at each school, focusing on the period when our SAT-takers were in high school (1997/98-2000/01) and an earlier period when they were in elementary school (1987/88-1990/91). We compute in each period the fraction black, fraction Hispanic, and fraction white at each public school. Data on racial composition for the later period are relatively complete, though we drop data without racial breakdowns from Tennessee in 1999/00 and 2000/01 and from Idaho in 1997/98, 1998/99, and 1999/00. In the earlier period, many states did not report schools' racial composition. When data are unavailable for any of the years 1987/88-1990/91, we substitute data from the earliest four year period for which data are available.[4]

We use these data to construct measures of segregation across high schools in the later period and across elementary schools in the earlier period. The later period is again the most straightforward. We multiply the school's total enrollment in grades 9-12—averaged over our four year window—by the racial shares to estimate the number of white and black high school students at each school. For private schools, for data are available only for 1997/98, we use that year's data. We then compute the average fraction black or Hispanic at schools attended by black and by white students, and compute our segregation measure as the difference between these.

To compute an elementary school segregation measure for the earlier period, we use average enrollment in grades 1-5 in 1988/91. For states for which we were forced to use later data to compute the racial composition, we use the same years to compute enrollment in our SAT-taker cohorts. That is, if we used data from 1991/92 through 1994/95, we total enrollment over grades 3-6 in 1991/92, 4-7 in 1992/93, 5-8 in 1993/94 and 6-9 in 1994/95. For private schools, however, we use grades 1-5 from the 1997/98 data in all states. We then again compute the number of white and black students in these grades at each school by multiplying the racial shares by the enrollment, average the fraction black or Hispanic over white and over black students' schools, and difference them.

[4] When data were available for at least two years of our four year window, we used the available data. In a few states that had data for 1990/91 but not for earlier years, we used data from 1991/92 as well.

Our final school segregation measure is a weighted average of the high school and elementary segregation measures, putting 2/3 weight on the latter to reflect the larger number of years spent in elementary school. We also create variants—e.g. using only public schools or using all grades from 1997/8-2000/1—to correspond to outcome variables that are measured accordingly. We have explored other definitions, and except as noted in the text results are generally insensitive to these decisions.

C. SAT data processing

The first step in our processing of the SAT data is to identify "SAT states," where the majority of college-bound students take the SAT exam rather than the alternative exam, the ACT. To do this, we compute the number of SAT takers in each state from our microdata sample, and define as SAT states those where the number of takers per year (averaged over our four cohorts) exceeds 25% of the 17-year-old population in the state in 2000. (A parallel analysis of ACT data results in an identical categorization.) We define a "SAT MSA" as one that is wholly contained within SAT states.

The next step is to create our background index. For this, we estimate separate regressions for white and black students' SAT scores, using only observations from SAT MSAs. These regressions include high school fixed effects and dummy variables for gender, 10 paternal education categories, 10 maternal education categories, and 14 family income categories. We take fitted values from these regressions, excluding the high school effects, as our background index.

Finally, we need a measure of the SAT-taking rate at each school. For most schools, we compute this as the ratio of the number of SAT-takers to the number of 12^{th} graders, each averaged over the four years in our data. Some schools appear in only a few years, and we use only the available years. A few schools, however, have very small 12^{th} grade enrollments but larger enrollments in earlier grades. When the number of SAT-takers exceeds the number of 12^{th} graders, we use 11^{th} (or 10^{th} or 9^{th}) grade enrollment for the denominator instead. We also compute test-taking rates for white and black students at the school, using the school white and black shares to compute the denominator. We use the resulting test-taking rates to construct the reweighting factors and inverse Mills ratios described in the Appendix.

III. Control Variables for Main Specification

Here we describe the construction of the control variables used in our primary specifications. All specifications in Table 3 include controls for the fraction black and the fraction Hispanic in the MSA's schools, averaged over the same years as are used for our school segregation measures. Specifications in Panel A also include the difference between black students' and white students' schools in the average inverse Mills ratio in the SAT participation rate, as described in Appendix B.

The "MSA demographic characteristics" introduced in columns B, E, and H of Table 3 are:

-log(population)
-Land area (in square miles)
-Fraction of the age-25+ population with a BA or more
-Fraction of the age-25+ population with some college
-The log of the mean household income in the MSA
-A Gini coefficient for household income in the MSA. As the household income distribution is reported in bins, we assume that each household is at the middle of its bin (i.e. those in the $15,000-20,000 bin are assigned $17,500), treating 0 as the bottom of the lowest bin (<$10,000) and assigning households in the ">$200,000" bin to the MSA mean income for households in that bin.

All of these are taken from MSA records on the 2000 Census SF3 file.

The "B-W background controls, SAT-takers" introduced in the same columns are:

-The black-white difference in the background index, discussed above
-The black-white difference in the fraction of SAT-takers' fathers with BAs
-The black-white difference in the fraction of SAT-takers' fathers with some college
-The black-white difference in the fraction of SAT-takers' mothers with BAs
-The black-white difference in the fraction of SAT-takers' mothers with some college
-The black-white difference in family incomes, on a SAT scale. Family incomes are reported in bins. These are converted to a SAT scale using a race-specific regressions of SAT scores on school fixed effects and dummies for each possible parental education and family income response. The family income measure is formed from the coefficients on the family income dummies. Students who did not respond to the parental education variables are assigned zero.

All of these are formed by averaging the reweighted SAT data to the MSA level separately for black and white students, then differencing.

Columns C, F, and I of Table 3 introduce controls for "B-W background controls, 0-17 year olds in Census data." These are:

-The black-white difference in the fraction of fathers with BAs
-The black-white difference in the fraction of fathers with some college
-The black-white difference in the fraction of mothers with BAs
-The black-white difference in the fraction of mothers with some college
-The black-white difference in the fraction of children living with one parent
-The black-white difference in the fraction of children living without either parent
-The black-white difference in the employment rate of children's mothers
-The black-white difference in the median family income of families with children
-The black-white difference in the fraction of children in poverty

The first four are computed from the PUMS data, and the remainder from the SF4 tabulation.

Finally, columns D, G, and J introduce black-white gaps in maternal and paternal residual wages. These are computed from our PUMS extract of the parents of children under age 17. For employed parents under age 65, we run race- and gender-specific regressions of log hourly wages—computed as annual wage and salary income divided by the product of annual weeks and weekly hours, and censored at $3 and $300 per hour—on MSA fixed effects, years of education, indicators for less than 12 years and 16+ years of education, and a cubic in potential experience. The black-white gap in maternal residual wages is the difference between the MSA fixed effects from the black mothers and white mother regressions, with a parallel definition for paternal residual wages.

IV. *Variables Used in Analysis of Changes in Segregation Over Time*

Two portions of our analysis study different aspects of changes in segregation over time. Figure 4 of the working paper (and the text describing it) examines whether shifting populations across MSAs has led to changes in the degree of segregation experienced by different groups (black or white, high or low education). For this analysis, we use only year-2000 measures of residential and school segregation, and examine changes in the proportions of various populations living in cities with different levels of 2000 segregation. Tables 5A and 5B of the working paper, by contrast, examine changes in the two segregation measures over time.

For Figure 4 of the working paper we use summary files from the Census in each year to compute the population distribution across MSAs. Our sources are the 1970 Census Summary Statistic File 4C (ICPSR study number 8107), the 1980 Summary Tape File 3A (ICPSR #8071), the 1990 Summary Tape File 3A (ICPSR #9782), and the 2000 Summary File 3 (ICPSR #13402). We use county-level population counts to form MSAs in each period's data according to the 1999 MSA definitions. MSAs in New England are therefore excluded from this analysis. When we distinguish between low- and high-education individuals of each race, we use thresholds that vary over time, attempting to maintain a constant fraction of each race's population in each category. The low education category is 7 or fewer years of education in 1970, 8 or fewer in 1980, and 11 or fewer in 1990 and 2000. The high education category is more than 12 years in 1970 and 16 or more years in all other years. The fraction of each race's population in each category is:

	1970	1980	1990	2000
Whites				
Low educ.	13.6%	16.6%	22.1%	16.4%
High educ.	22.4%	17.1%	21.5%	26.1%
Blacks				
Low educ.	33.0%	26.9%	36.9%	27.7%
High educ.	10.3%	8.4%	11.4%	14.3%

For Table 5A of the working paper, we use Cutler, Glaeser, and Vigdor's (1999) isolation indices from 1950-2000. The Cutler-Glaeser-Vigdor data are based on varying

city definitions. As a result, some of their cities and MSAs no longer exist in our 1999 definitions, some 1999 MSAs do not appear in their data, and some areas are split or combined between the two years. We aggregate to 1999 MSAs as best as the Cutler-Glaeser-Vigdor data allow, using the simple average when several of their units are combined into a single 1999 MSA. The first row of Table 5A shows the number of 1999 MSAs that can be observed in each year in the Cutler-Glaeser-Vigdor data.

For Table 5B of the working paper, we use the methods described above to compute high school and elementary level school segregation measures for both our early and our late periods. For the early period measures, in contrast to our lifetime measure, we exclude states where race data are unavailable in 1987/88-1990/91.

V. Construction of Desegregation Instrument

In Table 5, we instrument for school segregation with a measure of the strength of local court-ordered desegregation. Our data on desegregation orders are taken from Welch and Light (1987). Welch and Light obtained desegregation plans for a sample of school districts – mainly larger urban districts. Plans are categorized by type (busing, magnet schools, etc.) and by whether they were "major" or "minor" plans. Their data files also include dissimilarity indices for the districts' schools in each year. For each district, we compute the change in the dissimilarity index in the years surrounding the implementation of each major plan, and sum this over all major plans in the district. We then assign districts to 1999 MSAs. When there is more than one Welch and Light district in the MSA, we take the enrollment-weighted average of the various districts' measures. We then multiply the impact of the plan on the districts' dissimilarity indices by the fraction of metropolitan enrollment that is in these districts to obtain a crude measure of the impact of court-ordered desegregation in the Welch and Light districts on overall metropolitan segregation.

VI. Teacher characteristics

In Columns C-F of Table 8 of the working paper, we examine the relationship between segregation and relative black exposure to several teacher characteristics. Our data on teacher characteristics are taken from the Schools and Staffing Survey (SASS). We use the confidential version of the 1999-2000 SASS to assign teachers to the CCD and PSS records for their schools, then compute white and black enrollment-weighted averages of the teacher characteristics for schools with teachers in the SASS sample. Our dependent variables in Table 8 are the black-white differences in the weighted means of the various characteristics, using the race-specific enrollment distributions in the MSA as weights.

References

Cutler, David M., Edward L. Glaeser, and Jacob L. Vigdor (1999). "The Rise and Decline of the American Ghetto." Journal of Political Economy **107**(3): 455-506.

Welch, Finis and Audrey Light (1987). New Evidence on School Desegregation. Washington, D.C., United States Commission on Civil Rights.