ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-804

An Improved Method for Estimating School-Level Characteristics from Census Data

Shirin A. Hashim
Harvard University

Thomas Kelley-Kemple
Harvard University

Mary E. Laski
Harvard University

We propose a new method for estimating school-level characteristics from publicly available census data. We use a school's location to impute its catchment area by aggregating the nearest n census block groups such that the number of school-aged children in those n block groups is just over the number of students enrolled in that school. We then weight census data by the number of school-aged children in the block-group to estimate school-level measures. We conduct several robustness checks to assess the quality of our estimates and find that our method is broadly successful in replicating known school-level characteristics and producing unbiased estimates for school-level income. This method expands the available set of school-level variables to the broader and richer set of characteristics measured in the census, which can then be used to conduct descriptive and observational research across a long time horizon.

VERSION: July 2023

Suggested citation: Hashim, Shirin A., Thomas Kelley-Kemple, and Mary E. Laski. (2023). An Improved Method for Estimating School-Level Characteristics from Census Data. (EdWorkingPaper: 23-804). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/chqx-av23

An Improved Method for Estimating School-Level Characteristics from Census Data

Shirin A. Hashim, Thomas Kelley-Kemple & Mary E. Laski

Harvard Graduate School of Education, Cambridge, Massachusetts, USA

April 2023

We thank Douglas Staiger and Thomas Kane for their support and advisement on this project. We also thank seminar participants at the Harvard Graduate School of Education and the 2022 Society for Research on Educational Effectiveness annual meeting for helpful comments and suggestions.

Shirin A. Hashim
Harvard Graduate School of Education
13 Appian Way
Cambridge, MA 02138
shirin_hashim@g.harvard.edu

Mary E. Laski
Harvard Graduate School of Education
13 Appian Way
Cambridge, MA 02138
marylaski@g.harvard.edu

Thomas Kelley-Kemple
Harvard Graduate School of Education
13 Appian Way
Cambridge, MA 02138
tkelleykemple@g.harvard.edu

Abstract

We propose a new method for estimating school-level characteristics from publicly available census data. We use a school's location to impute its catchment area by aggregating the nearest *n* census block groups such that the number of school-aged children in those *n* block groups is just over the number of students enrolled in that school. We then weight census data by the number of school-aged children in the block-group to estimate school-level measures. We conduct several robustness checks to assess the quality of our estimates and find that our method is broadly successful in replicating known school-level characteristics and producing unbiased estimates for school-level income. This method expands the available set of school-level variables to the broader and richer set of characteristics measured in the census, which can then be used to conduct descriptive and observational research across a long time horizon.

I. Introduction

The recent publication of district-, and now school-level, achievement data on a common scale has revitalized interest in conducting analyses on a national scope (Fahle & Reardon, 2017; Reardon et al., 2021). However, researchers are limited in their ability to examine national trends as they relate to critical out-of-school factors due to a lack of administrative data. For example, while there is extensive evidence that socioeconomic status is a significant predictor of student outcomes, student-level administrative datasets usually only include a crude, binary measure of free- or reduced-price lunch (FRPL) eligibility (Owens, Reardon, & Jencks, 2016). Even when examining whole school characteristics, researchers must rely on aggregates of the limited student-level data or on census information about the census tract in which a school is located. Yet, attributing characteristics based on the tract of a school tends to bias estimates, as students are often drawn from a wider attendance area than these geographies capture (Saporito, 2017).

To estimate school-level characteristics using the census, the ideal data would include information about school attendance boundaries, sometimes known as school catchment areas. To date, the most comprehensive national dataset remains the National Center for Education Statistics' (NCES) School Attendance Boundary Survey (SABS) of 2015-2016. However, SABS is limited by the schools that responded to the survey and by the fact that further data collection has ceased. These data are insufficient for analyses that hope to examine national trends over long periods of time.

To combat this issue, we present the Census-to-School (CTS) method, which employs publicly-available census data and the public school universe data from NCES's Common Core of Data (CCD), to approximate school catchment areas and then estimate school-level characteristics for all census years. In brief, we use a school's location to impute its catchment area by aggregating the nearest *n* census block groups such that the number of school-aged children in those *n* block groups is roughly equal to the number of students enrolled in that school. We then weight census data by the number of school-aged children in the block-group to estimate school-level measures. This method expands the available set of school-level variables to a broader and richer set of characteristics, including average family income, rent costs, and parental education. These estimates, in turn, can be used to assess how school-level factors are associated with student achievement or to understand how school contexts moderate the effectiveness of policies or programs over time.

In this paper, we demonstrate the need for additional school-level data, describe the CTS method of estimating school-level data in detail, and evaluate the quality of CTS estimates through a variety of sensitivity checks. We first test whether we are capturing the "correct" students for each school by comparing to student zip codes from a nationally-representative sample. We find that the CTS method accurately captures the great majority of students in this sample; roughly 90% of the zip codes tied to students in NAEP data are identified through our method. We then test the accuracy of the method itself by checking our ability to reproduce existing school-level measures. We find that the CTS method performs particularly well in replicating existing administrative data across multiple metrics. Importantly, the CTS methods' accuracy is relatively consistent over the entire panel we study, which suggests our method can be used to account for historical trends in school qualities. We then use our method to estimate school-level income and test the accuracy of these estimates with existing data on school boundaries and school incomes. In the most conservative estimate, we find that our income estimates are generally plausible but somewhat overestimated. Finally, we compare our method to a simple “point-in-place” approach using schools’ census tracts and again find school-level estimates from the CTS method are more accurate, and this is consistent across our entire panel. Together, these checks suggest that the CTS method produces unbiased predictors of school-level measures and can be used by researchers to conduct descriptive and observational research with a wide range of school-level characteristics, and across a long time horizon.

II. Background

Scholars have argued that administrative geographies, such as census tracts or zip codes, are not interchangeable with school attendance boundaries (Saporito, Van Riper, and Wakchaure, 2013). However, in the absence of school boundary data, researchers often rely on these pre-defined geographies as convenient proxies when they need to make estimates about schools. For example, because schools rarely collect information on students’ digital connectivity, Barrett and Gerstenfeld (2020) use tract-level data from the American Community Survey (ACS) for the census tract in which a school is located in order to estimate how many students in a school had access to devices at home during pandemic-related school closures. Similarly, in the absence of students’ household income data, An (2013) and Persky et al. (1998) rely on the median income of a school’s zip code and a school’s census tract, respectively, to examine income-based gaps in health and education outcomes.

However, these “point-in-place” approaches pose many challenges (Saporito et al., 2007). Most importantly, the demographic composition of the area in which a school is located may not be reflective of the demographic composition of the school’s attendance boundary, let alone of the students who actually attend that school. For example, Saporito, Van Riper, and Wakchaure (2013) found that the correlation between the percent of non-Hispanic Black people in the census tract of a school and in the 7th grade school boundary was only 0.41 in 2011. Depending on the density of the population, residential segregation, school district redlining, or other political or geographic patterns, this relationship may also not be comparable over time or across districts.

While having school boundaries could alleviate some of these concerns, these data have rarely been collected at a large scale. In 2008, the National Science Foundation (NSF) funded a one-time creation of the School Attendance Boundary Information Systems database (SABINS) for the 2009-10 school year, which included boundaries from three states, over 400 of the largest school districts, and about 600 districts from 13 regionally diverse metropolitan areas (Saporito, Van Riper, and Wakchaure, 2013). Importantly, NCES worked with the U.S. Census Bureau to assign socio-demographic data from the ACS to each SABINS school attendance area, which created estimates for a new, rich set of school-level variables that were previously unavailable. However, Figure 1, which shows the school districts covered by SABINS, highlights that these data are far from nationally-representative, covering only about 15% of the nation's school districts.

To collect nationwide school attendance boundaries, NCES and the Census Bureau launched the School Attendance Boundary Survey (SABS) in 2013. This ultimately covered more than 70,000 schools across over 90% of the U.S. school districts in the 2015-16 school year, its most recent administration (see Figure 2). While these data are more geographically comprehensive than SABINS, they are not linked to census geographies to allow for any direct tabulation of school-level variables. To assign census data to schools using these boundaries, researchers must employ some method of assigning data across overlapping geographies.

In an extensive analysis of several of these data interpolation methods, Saporito et al. (2007) found that the "population weighting" method was most accurate in reproducing administrative data in the case of school-level variables. Unlike point-in-place approaches, this method does not assume that population characteristics are distributed evenly across different areas. Instead, it weighs the characteristics of larger units by actual block-level population counts. For example, to assign block group data to school attendance boundaries, one would overlay a map of school attendance boundaries on a map of block groups, and then assign block group characteristics to the school based on which block groups fall within the boundaries, weighted by the number of school age children in the block group. It is important to note, however, that the population weighting method still relies on school attendance boundary data, and these data are only available for a few years through SABINS and SABS. Researchers who want to look at trends in school-level variables cannot reasonably assume that the set of schools or their boundaries are necessarily consistent overtime.

For education researchers and policymakers, the lack of certain longitudinal school-level data becomes especially apparent when trying to examine the connection between a school's socioeconomic composition and student learning overtime. Despite the centrality of this relationship within education research (e.g., Kahlenberg, 2013; Jang & Reardon, 2019), school-level measures of students' socioeconomic status are imprecisely measured, most often proxied for by the share of students eligible for FRPL through the National School Lunch Program (NSLP). Although using FRPL eligibility is particularly convenient since it does not rely on having school attendance boundaries, and the data are easily accessible nationally and over time, there are widely-known issues with using it as a proxy for poverty. The dichotomous indicator does not align well with data on students' family income (Domina et al., 2018) and recent policy changes to how FRPL eligibility is determined have made it an inconsistent measure over time (Chingos, 2016; Greenberg, 2018). While several researchers and policymakers have recently proposed alternative ways to measure school poverty levels (e.g., Michelmore and Dynarski, 2017; Greenberg, 2018; Gutierrez, Blagg, and Chingos, 2022), these more reliable measures are only available for a limited number of years and rarely provide school-level information beyond poverty rates, such as average household income.

The CTS method addresses this gap. We build on the basic assumptions and strengths of the population weighting approach described above, but we do not rely on existing school attendance boundary data to determine which block groups should be included for each school, and we are not limited to describing socio-economic status solely in relation to the poverty line. Thus, this method provides a way for researchers to leverage existing census data to tabulate a vast set of school-level characteristics for public schools nationwide and over time.

III. Data

A. National Center for Education Statistics. We utilize several sources of data from the National Center for Education Statistics (NCES) for school-level information. Specifically, we use NCES's Common Core of Data (CCD) to get schools' locations by latitude and longitude[1], student enrollments by grade-level, and various school-level characteristics, including urbanicity, the number of students eligible for free- or reduced-price lunch, enrollment counts by race/ethnicity, and total enrollment. To conduct our verification checks of the method, we also use NCES's School Attendance Boundary Survey (SABS) to collect schools' boundary information for the 2015-16 school year, as well as data collected by the School Attendance Boundary Information Systems (SABINS) project for school attendance boundaries for the 2009-10 school year. Additionally, we use school administrative records and student questionnaire responses collected by NCES's National Assessment of Educational Progress (NAEP) to obtain home zip codes for a nationally representative sample of 8th grade students in the 2012-13 and 2014-15 school years. Lastly, we use family income data collected from the parents

[1] The CCD does not report school locations by latitude and longitude prior to the 2000-01 school year. For schools that we did not have a latitude-longitude location from a later year or for any schools that appeared to have moved, we used the school's mailing address (prior to 1999) or location address (1999-2000), along with the geocodehere Stata program (Hess, 2015), to obtain its latitude and longitude.

of a nationally representative sample of nearly 25,000 8th graders by NCES's National Education Longitudinal Survey of 1988 (NELS:88).

B. U.S. Census Bureau. For demographic data by census block groups and by census tracts, we use the decennial censuses and, starting in 2009, the American Community Survey (ACS). At the block group-level, we collect the number of children enrolled in school.[2] For each census tract, we collect the number of households with children in each income bin, the number of children by household type, the median income, and the number of individuals by age, race, and whether their family income is above or below the federal poverty level. Additionally, we use the U.S. Census Bureau's TIGER/Line Shapefiles to obtain block group boundaries in order to identify the geographic centroid of each block group.[3]

IV. CTS Method

To calculate school-level characteristics using the CTS method, we first estimate school catchment areas using schools' geographical coordinates and grade-level enrollments from the CCD. We rank the nearby census block groups by their geodetic distance from each school and identify the number of school-aged children in each block group from the census data. Starting from the block group nearest to each school, we collect successive block groups until the cumulative sum of school-aged children in the collected block groups just surpasses the school's enrollment. These collected block groups comprise our estimated attendance area of the school. In other words, if N_k is the number of school-aged children in block group k according to the census, and N_j is the number of

[2] The 1990 census only reports the number of children enrolled in school in grades 1-12. In later years, enrollment is reported for grades 1-4, 5-8, and 9-12 separately. For 1990, we assume that enrollment is equal in each grade band.

[3] We use the shp2dta (Crow, 2006) program in Stata to identify the centroid of each census block group from the TIGER/Line Shapefiles.

students enrolled in school j according to the CCD, then we associate the nearest K block groups that satisfy the inequality:

$$\sum_{k=1}^{K-1} N_k < N_j \leq \sum_{k=1}^{K} N_k$$

to each school j, where block groups k are ordered in increasing distance from school j's location.

Because most data regarding household characteristics are censored at the block group level, we assign tract-level census characteristics to each block group identified as being in each school's estimated catchment area. To estimate a school-level characteristic, we then take the average of the tract-level characteristics for each block group assigned to a school, weighted by the number of school-aged children in each block group. That is, we estimate that the proportion of students in school j who identify as some characteristic i is given by:

$$P_{ji} = \frac{\sum_{k=1}^{K} p_{ki} N_k}{\sum_{k=1}^{K} N_k}$$

where p_{ki} is the proportion of students in block group k's corresponding census tract who identify as some characteristic i. We can implement this process using decennial censuses (and the ACS after 2009). For intervening years, we use a linear interpolation of the school-level measures from the two nearest decennial censuses.

In the next section, we evaluate the quality of CTS estimates through a variety of sensitivity checks. We first test whether we are capturing the "correct" students for each school by comparing to student zip codes from a nationally-representative sample. We then test the accuracy of the method itself by checking our ability to reproduce existing school-level measures. We then use our method to estimate school-level income and test the accuracy of these estimates with existing data on school boundaries and school incomes. Finally, we compare our method to a simple "point-in-place" approach using schools' census tracts. Across all of our checks, we find that the CTS method produces reasonably accurate and unbiased predictors of school-level measures. Our checks suggest that the CTS method can be used by researchers to conduct descriptive and observational research with a wide range of school-level characteristics across a long time horizon.

V. Verification Checks

A. Capturing Students in Schools. The ideal dataset needed to assess whether the CTS method is accurately capturing the students in each school would include students' home addresses and their school enrollment. We could then see if our estimated school catchment areas are correctly assigning the home addresses of the students to their correct school. However, these data have never been captured at a national scale nor over an extended period of time. Instead, we turn to the NAEP, which collected a nationally-representative sample of 8th graders' home zip codes in the 2012-13 and 2014-15 school years. These data come from two sources: administrative records provided to NAEP by each school and NAEP's student questionnaires, in which students were asked to write the zip code of their home address in an optional free-form field. In each year, we have at least one source of home zip code for over 95% of all 8th grade students who took the NAEP, along with which schools the students were attending.

Because the CTS method assigns census block groups to each school, not zip codes, we first identify the census tract associated with each block group and use a tract-to-zip code crosswalk[4] to get the CTS-derived catchment areas[5] in terms of zip codes. Then, for each student in the NAEP, we are able to see whether their school's CTS-derived catchment area correctly captures their zip code. Table

[4] The U.S. Department of Housing and Urban Development's (HUD) Office of Policy Development and Research (PD&R) releases zip code crosswalk files quarterly each year. Here, we use the 1st quarter crosswalk for 2013 and 2015, obtained from: https://www.huduser.gov/portal/datasets/usps_crosswalk.html. Each census tract maps to 1 to 51 zip codes, with both a median and average of approximately 2 in each year.

[5] Appendix A includes a test for the accuracy of these CTS-derived school catchment areas compared to SABINS and SABS.

1 summarizes these findings. For both years in which we have an approximation of school catchment areas from the NAEP, we find that 88% of students' home zip codes are correctly assigned by the CTS method. Assuming there is considerable noise in the student-reported zip code data from the NAEP, we also conduct a similar test using only students with administrative zip codes.[6] As shown in Column 2, limiting to administrative NAEP records, we find that 89% and 90% of students' home locations are captured by the CTS method for 2013 and 2015, respectively. Alternatively, in Column 3, we limit our sample to students whose home zip code appeared more than once in a given school, and find that 93% and 94% of students' home locations are captured by the CTS method each year. Thus, it seems the method is largely able to allocate the correct students to each school. This suggests that the approach could produce reasonable estimates of school-level characteristics, which we test directly in the next section.

B. School-Level Measures. To evaluate the accuracy of the CTS estimates of school characteristics, we reproduce school characteristics that are available in administrative data. Specifically, we compare school-level characteristics available in the CCD to their comparable CTS estimates in three ways. We first present the correlation between the CCD values and the CTS estimates. We then look at the average difference between the CCD and CTS values as a measure of bias. Finally, we report the square root of the average squared difference (the root mean square deviation, or RMSD), which aggregates the magnitudes of the errors in our estimates into a single measure of accuracy.[7] These findings are presented in Table 2. We find that the CTS method performs

---

[6] To approximate the degree of noise in the student-reported data, we can look at the 81,260 and 68,850 students who had both sources of home zip code in 2013 and 2015, respectively. Of these students, 85% and 84% had matching student-reported and administrative home zip code records, suggesting considerable noise in the student-reported data.

[7] The RMSD is calculated as: RMSD = $\mathrm{sqrt}(\sum(CCD_j - CTS_j)^2)$, where CCD_j is the administrative characteristic for school j, and CTS_j is the CTS estimate of that characteristic. The RMSD is a weighted measure of accuracy of the estimate given on the same scale as the target characteristic. The closer the RMSD is to 0, the more accurate the estimates are, on average. In this instance, the RMSD is on a scale of 0 to 1 and can be interpreted

quite well in replicating administrative data across all of these metrics: correlations are all moderate to strong, and average differences and RMSDs are generally close to zero. Importantly, these relationships are consistent over the entire panel we study; we find no evidence that the bias of our estimates is notably changing over time.

We first compare CTS estimates of poverty rates with the proportion of students eligible for FRPL at a given school, shown in Panel A of Table 2. To approximate FRPL rate in the CTS method, we calculate the proportion of households assigned to each school that fall below a designated "FRPL line" for each year.[8] Across all studied years, the correlation between the CTS estimate and the administrative data remains close to 0.7.[9] Importantly, we do not find any evidence that the reliability of our method has decreased over time: estimates are consistent across the entire panel.[10] The average difference between schools' FRPL rates in the CCD and our estimates of FRPL rates in the CTS method shows that while our method underestimates the FRPL rate, on average, the level of bias in our estimate is stable over time. Similarly, the RMSD, or the weighted average error between the CTS estimate of FRPL rates and actual FRPL rates in the CCD is consistently around 0.2.

Our method performs even better at approximating the racial breakdown of schools. When comparing the proportion of students in a school who are Black or White to the CTS racial estimates,

---

as the average error that the CTS estimates have when compared to the actual school characteristics found in the CCD, where more weight is given to larger estimation errors.

[8] In general, the FRPL eligibility is based on a student's family income being below 185% of the federal poverty level. To calculate an estimated "FRPL line" for each year, we use data on annual poverty thresholds from the US Census. For each year, we take the weighted average poverty threshold for four-person families and multiply by 1.85. Poverty estimates can be found here: https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-thresholds.html.

[9] 52% of schools were missing FRPL data in the 1990 CCD, so this analysis is limited to 2000-2015.

[10] While our method's strong correlation with CCD FRPL rates is encouraging, we acknowledge that FRPL is an imperfect, flawed measure of school-level income. The Urban Institute recently created a more sophisticated measure of school poverty (Gutierrez, Blagg, and Chingos, 2022). This measure is only available for the years 2013 through 2018. When we compare our CTS FRPL estimates to the Urban Institute poverty measures, we again find strong correlations ranging from 0.75 to 0.8. We find this to be further confirmation that our method is accurately capturing school-level income.

correlations are consistently above 0.8 between 1990 and 2015. The average differences and the RMSDs are also stable across all years and are closer to 0, suggesting that the CTS estimates of schools' racial makeups are relatively more accurate, on average. Together, these measures imply that, while the CTS estimates are subject to some error, there is no evidence that the reliability of our estimates is changing over time.[11]

C. Mean and Variance of School-Level Income. Given the demonstrated evidence of the importance of socioeconomic status in predicting student outcomes (Reardon, 2011), we believe that one of the most important applications of the CTS method is approximating more nuanced information about school income. As such, in this section, we first describe how we estimate school-level mean income and variance of income and then assess how well the CTS method is able to estimate these measures.

Estimating School-Level Income

Recall that the CTS method is estimating the proportion of students in a school who identify as some characteristic that is available in the census data at the tract-level. To report information about household income, the census organizes incomes into a set of ranges or bins and reports the number of families that fall into each. Thus, if we apply the CTS method to each of those income bins in the census, we are able to estimate the proportion of students in a school that fall into each income bin.

In order to estimate a school's mean income and variance of income from these bucketed data, we first start by assuming that the log of income is distributed normally for each school.[12] We then fit a log-normal distribution to the set of income bins for each school, an approach outlined by vonHippel,

[11] Appendix B provides RMSDs between administrative school characteristics and CTS estimates for different school types. Broadly, we find that our method performs slightly better in more suburban areas, but still works reasonably well for urban schools, and that the RMSDs are similar and consistent over time when computed separately for primary and middle schools. Overall, we find no evidence that our measure is more or less biased for certain schools based on urbancity and level.

[12] Log transformation of income is common given the right skew of the income distribution.

Scarpino, and Drown (2017). If the log of income, $ln(Y)$, is distributed normally, and C_l (expressed in log 2016 dollars) is the upper range of the lth income bin (where l=1,...,15 corresponding to 16 income bins), then the cumulative proportion of the school enrollment with income below C_l can be expressed as:

$$P_{jC_l} = \Phi\left(\frac{\ln(C_l) - \mu_{Y_j}}{\sigma_{Y_j}}\right)$$

where μ_{Y_j} and σ_{Y_j} represent the mean and standard deviation of log income in school j. Rearranging terms from this equation results in the following expression, which is fit for each school:

$$ln(C_l) = \mu_{Y_j} + \sigma_{Y_j} invnorm\left(P_{jC_l}\right) + \vartheta_{jl}$$

In other words, we fit a linear relationship between the upper bound of each of the income bins and the inverse normal of our estimate of the cumulative proportion of students below that upper bound for each school. The intercept of the resulting best fit line is then our estimate of the mean income for the school, and the squared slope of that line is our estimate of the variance of income.

Income Verification Checks

We test these income estimates in two distinct ways. First, we estimate the mean and variance of income using the same approach described above but with administrative school boundary data from SABINS. This allows us to compare the quality of the boundary estimates while keeping the quality of information regarding income constant. Table 3 presents correlation and regression coefficients comparing our estimated income means and variances using these two different sets of school boundaries. We find that our income estimates using the CTS boundaries align well with the same estimates using the true school attendance boundaries. The correlations between the estimates are quite strong ($r > 0.7$). The regression coefficients are also very close to one, indicating that our CTS method does not create a school boundary that encompasses a substantially different student population than the true catchment area. Figures 3 and 4 display these estimates in more detail, with each point on the graphs representing one school. These figures provide further evidence that, in general, our CTS estimates align well with the estimates using the SABINS boundaries.

Our second verification test employs a more precise measure of average school income than one inferred based on administrative school boundaries. Here, we use the NELS:88 base-year survey, which collected parent-reported family income data of participating 8th grade students from a nationally representative sample of over 1,000 schools. Schools' mean and variance of income are estimated using the 25-30 students who were sampled from each school and incorporate weights to account for the probability of selection and non-response. NELS is the most appropriate, and perhaps the only, survey that can provide reasonable estimates for school-level mean income that we can compare against our CTS estimates. That is because it is nationally-representative at the school-level, covers a time period that we are able to estimate income using the CTS method, and collects individual-level family income from middle school students.

Again, the NELS income data were reported in binned categories, so we apply the statistical approach outlined previously to infer a school-level mean income based on the weighted proportion of students in each income bin. Figure 5 compares the resulting estimates of school log income for 1987 based on the NELS data to the CTS estimates from the same year. If the CTS estimates were perfectly aligned with the NELS estimates, all of the school observations would fall along the bolded Y=X line. Instead, we see that the CTS estimates tend to be a bit higher than the survey mean income estimates for schools with lower mean income, and the CTS estimates tend to be a bit lower than the survey mean income estimates for schools with higher mean income. Still, it appears the CTS method does reasonably well in estimating schools' mean income, further evidenced by the high correlation between the survey estimates and the CTS estimates (r=.767). Figure 5 additionally shows that rural schools (blue) tend to have higher mean income than suburban (orange) and urban (gray) schools, but it does not seem that the CTS method is doing systematically better or worse for any one type of school.

Beyond the visual comparison, we can test the differences between the NELS and CTS estimates of mean income more formally using a series of null-hypothesis tests. That is, if we assume that the CTS estimate of a school's mean income is in fact its true mean (μ), then we can conduct a null-hypothesis test to obtain the likelihood that we would observe a NELS sample mean ($\bar{x}$) that is equal to or more extreme than the one we estimate based on the NELS under the null distribution. In other words, we calculate the following t-statistic for each school:

$$t_j = \frac{\overline{x_j} - \mu_j}{s_{x_j}/\sqrt{n_j}}$$

where s_{x_j} is the standard error of students' family income in the NELS and n_j is the number of students sampled in the NELS. This test of our method is the most conservative test we can conduct because we are assuming that there is no error in the CTS estimates when we use them as the true means in estimating the t-statistic for each school. In actuality, the CTS estimates are also based on sampling data from the census, so there is innate uncertainty in the estimates that is not being taken into account. As such, the findings below should be interpreted as the upper bound of the potential error in our CTS school income estimates, not a precise measure of the quality of our estimates.

Figure 6 plots the resulting t-statistic for each school against the NELS sample mean income. The red lines indicate the critical values for the 95% confidence interval. Based on a two-sided test, we would expect about 10% of our schools to fall outside these critical values by chance or 5% on each side. The dots outside of these critical value lines are schools for which it seems unlikely that the survey estimate for mean income could have been drawn from a distribution of sample means centered around our CTS estimate of mean income for the school. We see that approximately 30% of schools fall outside of these lines, rather than the expected 10%, and that high income schools more likely to have their income underestimated. At the same time, there does not seem to be a trend in how the method performs based on urbanicity.

We can also conceptualize the results of the null hypothesis tests in terms of the corresponding p-values, which tell us the plausibility that the NELS sample mean of school income could have been drawn from a null distribution centered around the CTS estimate. Figure 7 shows the distribution of p-values for all of the schools in the NELS survey data for which we have CTS estimates of school mean income. Under the null distribution, we would expect a uniform distribution with approximately 5% of schools falling into each of the 5 percentage point ranges of p-values as indicated by the red horizontal line. Again, we see that there is an unexpected proportion of schools with survey mean income estimates that are extreme. That is, about 25% of schools have an estimated p-value between .95 and 1, which is 20 percentage points higher than we would expect by chance. In other words, the NELS estimate for school mean income and the CTS method's estimate are not reasonably similar for about 1 in 5 schools. However, it is important to note again that this is the most conservative test of the method. Thus, we argue that our income estimates are generally plausible, albeit possibly overestimated for up to 20% of schools.

D. Comparing CTS to Point-in-Place. With complex methods such as the CTS, it is important to consider the tradeoff between simplicity and accuracy. As described in Section II, researchers often use simple point-in-place approaches to proxy for school income (An, 2013; Persky et al., 1998). These methods use characteristics of a school's geography, such as median income in a school's zip code or census tract, as a "good enough" estimate of income. These estimates are clearly biased (Saporito, 2017), but the magnitude of the error with this approach is unclear. In this section, we compare the accuracy of this type of point-in-place approach to the CTS method in recreating school-level characteristics.

To test this, we use data on schools' racial breakdown by year from the CCD. For brevity, we focus here on the proportion of students in a school who are Black or White. We then make two comparisons: we compare a school's true, CCD-reported proportion of Black (or White) students in a given year to (1) the proportion of children in the school's census tract that are Black (or White), and (2) the proportion of children in our CTS method boundary that are Black (or White). The first comparison represents the characteristic we would estimate if using the point-in-place approach.

Figures 8 and 9 present the results of these comparisons. In both figures, black lines with dots represent the relative proportions of Black students and gray lines with diamonds represent the relative proportions of White students. The solid lines represent the traditional point-in-place method using the census tract of a school and the dashed lines represent our CTS method. When comparing these two methods to CCD school-level characteristics, we consistently find that the CTS outperforms the simple point-in-place method across all comparisons and years. In Figure 8, we plot the raw differences between the CCD proportion Black (or White) and the estimated proportion from each of the two methods. The difference between the CCD and our method is always closer to zero than the difference between the CCD and the point-in-place method. While the relative differences may appear small, the difference between the CTS method and CCD is consistently at least 20% smaller than the difference between the point-in-place method and CCD.

Figure 9 presents the square root of the average squared difference (root mean square deviations, or RMSDs) comparing the CCD proportions to the CTS method (dashed) and the point-in-place method using census tract (solid). Again, the CTS estimates are consistently better than the point-in-place method; our RMSDs are always closer to zero and indeed seem to be improving over time. We argue that these simple comparisons are convincing evidence that the CTS method outperforms simple point-in-place proxies consistently and meaningfully.

VI. Conclusion

The CTS method expands the number of school-level variables that can be estimated using publicly available census data. In this paper, we focus on school-level household income estimates, which are a critical component for understanding how income-based achievement gaps have changed over time (Hashim et al., 2020), but census data include a wide range of information including reports on household possessions, professional status of household members, and housing costs. The CTS method, and our corresponding publicly available dataset[13], may assist future researchers to answer a broader range of educational research questions, bounding causal effects across schools (Miratrix et al., 2017) and generalizing to wider populations of schools (Tipton & Olsen, 2018).

One important consideration is that the CTS estimates will innately reflect some degree of measurement error. As such, we strongly discourage the use of these estimates to describe an individual school or subset of schools. While we have demonstrated that, collectively, these estimates are unbiased, for most schools in most years, we have no way of assessing the quality of an individual school's estimate. The enterprise of testing new methods for estimating school-level characteristics is hampered by the very lack of data that we aim to address. Still, our checks show alignment between the limited administrative data that do exist and our CTS estimates of those data. Moreover, we show consistency in that alignment over almost 30 years in spite of the fact that open-enrollment policies, charter schools, and other non-residence based forms of school attendance have increased during this period. As a result, we are optimistic about the application of these data for exploratory, historical, and observational research.

[13] Data linking census block groups to schools using the CTS method, as well as school-level mean income and variance estimates are openly available in Harvard Dataverse at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/G3PBNZ.

VII. References

An, B. P. (2013). The Impact of Dual Enrollment on College Degree Attainment: Do Low-SES Students Benefit? *Educational Evaluation and Policy Analysis*, 35(1), 57–75.

Barrett, N. & Adam Gerstenfeld (2020). Closing the Digital Divide. What Policymakers Need to Know. COVID-19 Research & Evaluation. *National Alliance for Public Charter Schools*.

Chingos, M. (2016). No More Free Lunch for Education Policymakers and Researchers. *Brookings Institution Reports.*

Crow, K. (2006). "SHP2DTA: Stata module to converts shape boundary files to Stata datasets," Statistical Software Components S456718, Boston College Department of Economics, revised 17 Jul 2015.

Domina, T., Pharris-Ciurej, N., Penner, A.M., Penner, E.K., Brummet, Q., Porter, S.R., and Sanabria, T. (2018). Is Free and Reduced-Price Lunch a Valid Measure of Educational Disadvantage? *Educational ResearcherI,* 47 (9): 539–55.

Fahle, E. & Reardon, S.F. (2018). How much do test scores vary among school districts? New estimates using population data, 2009-2015. *Educational Researcher*, 47(4).

Geverdt, D., (2018). *School Attendance Boundary Survey (SABS) File Documentation: 2015-16* (NCES 2018-099). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Greenberg, E. (2018). New Measures of Student Poverty: Replacing Free and Reduced-Price Lunch Status Based on Household Forms with Direct Certification. *Urban Institute.*

Gutierrez, E., Blagg, K., & Chingos, M. (2022). Model Estimates of Poverty in Schools. *Urban Institute.*

Hashim, S.A., Kane, T.J., Kelley-Kemple, T., Laski, M.E., & Staiger, D.O. (2020) Have Income-Based Achievement Gaps Widened or Narrowed? *NBER Working Paper*, 27714.

Hess, S. (201). "GEOCODEHERE: Stata module to provide geocoding relying on Nokia's Here Maps API," Statistical Software Components S458048, Boston College Department of Economics, revised 21 May 2020.

Jang, H. & Reardon, S.F. (2019) States as Sites of Educational (In)Equality: State Contexts and the Socioeconomic Achievement Gradient. *AERA Open*, 5(3).

Kahlenberg, R.D. (2013). From All Walks of Life: New Hope for School Integration. *American Educator*, 36(4), 2.

Manson, S., Schroeder, J., Van Riper, D., & Ruggles, S. (2019) IPUMS National Historical Geographic Information System: Version 14.0. Minneapolis, MN: IPUMS.

Michelmore, K. & Dynarski. S. (2017). The Gap within the Gap: Using Longitudinal Data to Understand Income Differences in Educational Outcomes. *AERA Open,* 3(1): 1–18.

Miratrix, L., Furey, J., Feller, A., Grindal, T., Page, L.C. (2017) Bounding, an Accessible Method for Estimating Principal Causal Effects, Examined and Explained. *Journal of Research on Educational Effectiveness*, 11(1), 133-162.

Owens, A., Reardon, S.F., & Jencks, C. (2016). Income Segregation Between Schools and School Districts. *American Educational Research Journal*, 53(4), 1159–1197.

Persky, V.W., Slezak J., Contreras, A., & Becker, L. (1998). Relationship of Race and Socioeconomic Status with Prevalence, Severity, and Symptoms of Asthma in Chicago School Children. *Annals of Allergy, Asthma, & Immunology*, 81(3):266.

Reardon, S.F. (2011) "The Widening Academic Achievement Gap Between the Rich and the Poor: New Evidence and Possible Explanations." in Greg J. Duncan and Richard J. Murnane (eds.) *Whither Opportunity?* New York: Russell Sage Foundation.

Reardon, S.F., Weathers, E.S., Fahle, E.M., Jang, H., & Kalogrides, D. (2021). Is Separate Still Unequal? New Evidence on School Segregation and Racial Academic Achievement Gaps.

Saporito, S. (2017). Shaping Income Segregation in Schools: The Role of School Attendance Zone Geography. *American Educational Research Journal*, 54(6), 1345–1377.

Saporito, S., Van Riper, D., & Wakchaure, A. (2013). Building the School Attendance Boundary Information System (SABINS): Collecting, Processing, and Modeling K to 12 Educational Geography. *J Urban Reg Inf Syst Assoc.*, 25(2), 49-62.

Saporito, S., Chavers, J.M., Nixon, L.C., & McQuiddy, M.R. (2007). From Here to There: Methods of Allocating Data Between Census Geography and Socially Meaningful Areas. *Social Science Research*, 36(3): 897–920.

The College of William and Mary and the Minnesota Population Center. (2011). School Attendance Boundary Information System (SABINS): Version 1.0. Minneapolis, MN: University of Minnesota.

Tipton, E. & Olsen, R.B. (2018). A Review of Statistical Methods for Generalizing From Evaluations of Educational Interventions. *Educational Researcher*, 47(8), 516–524.

United States Department of Education. National Center for Education Statistics. National Education Longitudinal Study, 1988.

vonHippel, P.T., Hunter, D.J., & Drown, M. (2017). Better estimates from binned incomes: Interpolated CDFs and mean-matching. *Sociological Science*, 4, 641-655.

VIII. Appendices

Appendix A: Estimating School Catchment Areas

To assess the accuracy of our school catchment areas, we compare our inferred attendance areas to the two available administrative datasets of nationwide school attendance areas: 2009-10 SABINS and 2015-16 SABS. In the SABINS, school boundaries were block-rectified. That is, true school boundaries were overlaid on census blocks and then redrawn to include all blocks with centroids that fall within the true boundary. When we compare the list of block groups in this block-rectified boundary set to the block groups included in our CTS method, we find that 47% of block groups and 51% of tracts assigned to a school in the SABINS are also assigned to the same school in the CTS method. When looking at the school-level, 51% of an individual school's SABINS-assigned block groups are also assigned in our CTS method, on average.

Recall that SABINS coverage is incomplete and includes only a third of the total schools we include in our CTS method or about 15% of school districts nationwide (see Figure 1). Thus, we also conduct this analysis with the SABS data, which has attendance boundary information for over 70,000 schools across 12,000 districts in the United States. While this dataset represents perhaps the most complete single source of data for examining attendance boundaries across both districts and states, our findings based on SABINS are nearly identical to those based on SABS. When we compare the list of block groups in in a SABS-provided attendance area to the set to the block groups included in our CTS method, we find that 47% of block groups and 53% of tracts assigned to a school in the SABS are also assigned to the same school in the CTS method. When looking at the school-level, 50% of an individual school's SABS-assigned block groups are also assigned in our CTS method, on average.

While there is some variability in the match rate between these administrative records and the CTS method based on school geography and demographics, it is clear that the CTS method is not particularly effective in reproducing school attendance boundaries themselves. However, this does not preclude it from reliably estimating school-level data. The CTS-inferred boundaries may still be capturing the students who attend each school or demographically-similar students, such that the school-level estimates are comparable.

Appendix B: Comparing School Characteristics in the CCD to Estimates using the CTS Method by School Urbanicity and Level

Appendix Table B1: RMSD Between School Characteristics in the CCD and Estimates using the CTS Method, by School Urbanicity

	1990	2000	2009	2011	2013	2015
A. *Percent of Students Receiving FRPL*						
Rural Schools	N.A.	0.166	0.211	0.215	0.221	0.226
Suburban Schools	N.A.	0.184	0.225	0.223	0.229	0.233
Town Schools	N.A.	0.168	0.201	0.203	0.209	0.215
Urban Schools	N.A.	0.245	0.285	0.277	0.267	0.274
B. *Proportion of Students Black (including Black-Hispanic)*						
Rural Schools	0.076	0.073	0.086	0.081	0.079	0.079
Suburban Schools	0.133	0.103	0.104	0.102	0.097	0.098
Town Schools	0.080	0.082	0.080	0.078	0.074	0.074
Urban Schools	0.187	0.164	0.156	0.151	0.146	0.144
C. *Proportion of Students White*						
Rural Schools	0.143	0.113	0.135	0.133	0.134	0.134
Suburban Schools	0.176	0.177	0.133	0.132	0.131	0.131
Town Schools	0.142	0.158	0.116	0.119	0.117	0.118
Urban Schools	0.241	0.243	0.170	0.162	0.158	0.157

Note: 52% schools are missing FRPL information in the CCD in 1990, 13% are missing in 2000, and less than 2% are missing in 2009 and beyond. The CCD and census handle race/ethnicity differently. For both datasets, the Black category includes both Black Hispanics and Black non-Hispanics. In the CCD, the White category always includes both White Hispanics and White non-Hispanics. The census codes both White Hispanics and White non-Hispanics as "White" in 1990 and 2000. From 2009 on, the CTS estimates include only White non-Hispanics but are still compared to the CCD proportions that include White Hispanics.

Source: U.S. Department of Education. National Center for Education Statistics. Common Core of Data: Public Elementary/Secondary School Universe Survey

Appendix Table B2: RMSD Between School Characteristics in the CCD and Estimates using the CTS Method, by School Level

		1990	2000	2009	2011	2013	2015
A.	*Percent of Students Receiving FRP*L						
	Primary Schools	N.A.	0.163	0.204	0.204	0.207	0.215
	Middle Schools	N.A.	0.191	0.237	0.232	0.235	0.242
B.	*Proportion of Students Black (including Black-Hispanic)*						
	Primary Schools	0.110	0.101	0.098	0.091	0.090	0.088
	Middle Schools	0.131	0.106	0.103	0.099	0.097	0.097
C.	*Proportion of Students White*						
	Primary Schools	0.174	0.167	0.121	0.116	0.113	0.114
	Middle Schools	0.180	0.180	0.137	0.131	0.130	0.131

Note: 52% schools are missing FRPL information in the CCD in 1990, 13% are missing in 2000, and less than 2% are missing in 2009 and beyond. The CCD and census handle race/ethnicity differently. For both datasets, the Black category includes both Black Hispanics and Black non-Hispanics. In the CCD, the White category always includes both White Hispanics and White non-Hispanics. The census codes both White Hispanics and White non-Hispanics as "White" in 1990 and 2000. From 2009 on, the CTS estimates include only White non-Hispanics but are still compared to the CCD proportions that include White Hispanics.

Source: U.S. Department of Education. National Center for Education Statistics. Common Core of Data: Public Elementary/Secondary School Universe Survey

IX. Tables and Figures

Table 1: Percent of Students' Home Zip Codes Captured by CTS-Derived School Catchment Areas

	All Students	Administrative Zip Code Records Only	Repeated Zip Codes Only
2012-2013			
Percent of Students Captured	88.3%	88.9%	93.2%
N	*156,970*	*94,890*	*144,750*
2014-2015			
Percent of Students Captured	88.3%	90.1%	93.8%
N	*126,420*	*79,910*	*115,560*

Note: The analysis sample in Column 1 accounts for 91% of the 8th grade students who took the NAEP mathematics assessment in each year. These students had at least one record of their home zip code (an administrative record and/or a student report) and attended a public school that was included in the CTE method. Column 2 excludes students with only a student-reported zip code, limiting the sample to administrative zip code records. Column 3 limits the sample to students whose zip codes were represented more than once in their school in a given year. All counts are rounded to the nearest 10, per IES reporting requirements. Percentages are calculated using NAEP's student-level weights.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2013 and 2015 Mathematics Student and School Questionnaires

Table 2: Comparing School Characteristics in the CCD to Estimates using the CTS Method

	1990	2000	2009	2011	2013	2015
A. *Percent of Students Receiving FRPL*						
Correlation between CCD and Estimate	N.A.	0.715	0.669	0.655	0.694	0.690
Average Diff. between CCD and Estimate	N.A.	0.034	0.112	0.100	0.127	0.134
RMSD	N.A.	0.194	0.235	0.234	0.235	0.241
N	–	*60,860*	*74,930*	*76,400*	*76,130*	*75,820*
B. *Proportion of Students Black (including Black-Hispanic)*						
Correlation between CCD and Estimate	0.853	0.903	0.900	0.899	0.901	0.898
Average Diff. between CCD and Estimate	0.005	0.020	0.021	0.013	0.010	0.010
RMSD	0.128	0.113	0.113	0.109	0.106	0.106
N	*59,470*	*67,970*	*76,610*	*76,770*	*76,630*	*76,560*
C. *Proportion of Students White*						
Correlation between CCD and Estimate	0.828	0.859	0.912	0.914	0.914	0.913
Average Diff. between CCD and Estimate	-0.045	-0.063	-0.026	-0.019	-0.020	-0.020
RMSD	0.179	0.183	0.142	0.139	0.138	0.138
N	*60,796*	*67,972*	*76,607*	*76,773*	*76,630*	*76,564*
Total # Schools in CCD	*68,300*	*69,790*	*76,610*	*76,780*	*76,630*	*76,560*

Note: 52% schools are missing FRPL information in the CCD in 1990, 13% are missing in 2000, and less than 2% are missing in 2009 and beyond. The CCD and census handle race/ethnicity differently. For both datasets, the Black category includes both Black Hispanics and Black non-Hispanics. In the CCD, the White category always includes both White Hispanics and White non-Hispanics. The census codes both White Hispanics and White non-Hispanics as "White" in 1990 and 2000. From 2009 on, the CTS estimates include only White non-Hispanics but are still compared to the CCD proportions that include White Hispanics.

Source: U.S. Department of Education. National Center for Education Statistics. Common Core of Data: Public Elementary/Secondary School Universe Survey

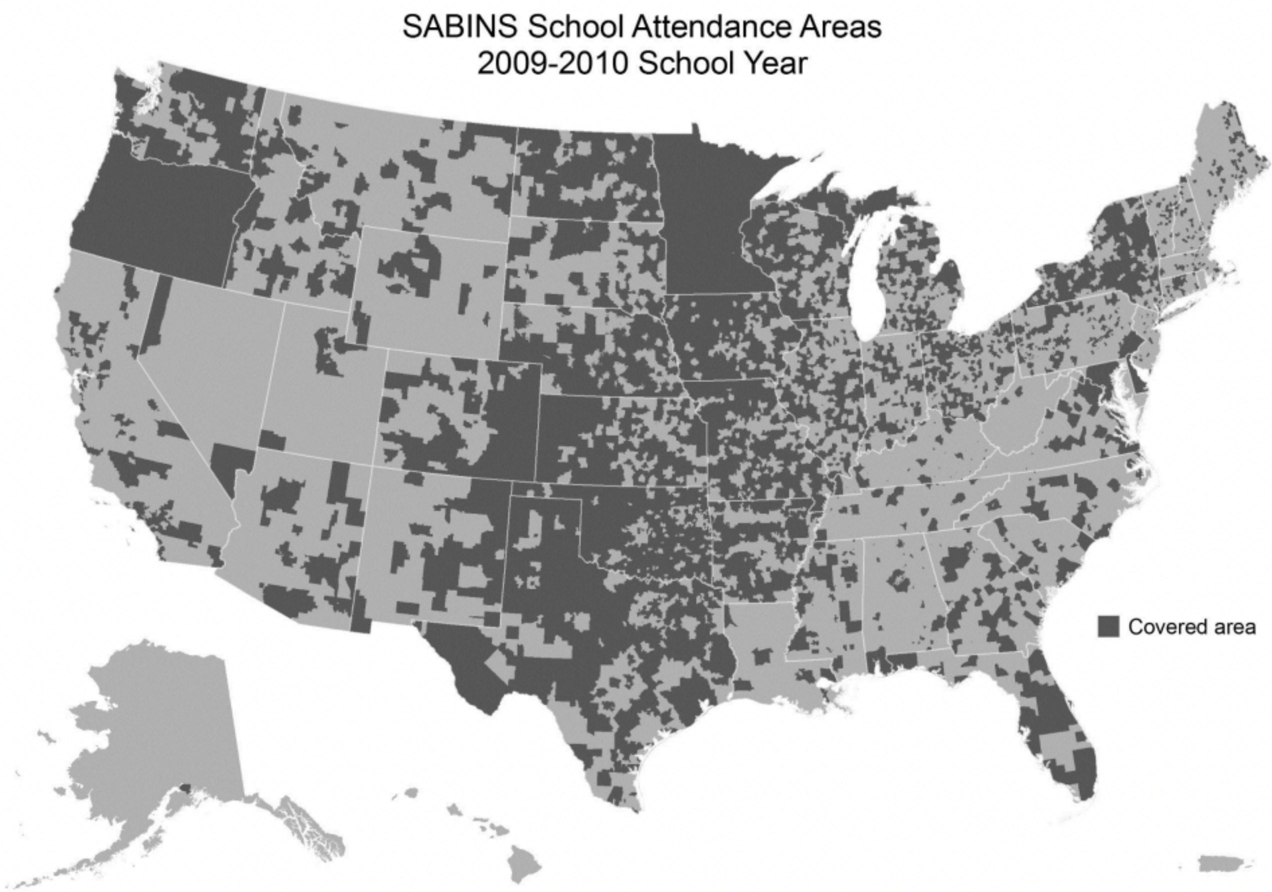
Table 3: Correlations Between Income Estimates using SABINS 2009 Boundaries and CTS Method using 2011 Census

Estimated Income Mean	
Correlation between SABINS and CTS	0.919
Regression Coefficient on CTS Estimate in Predicting SABINS	0.959
N	*29,420*
Estimated Income Variance	
Correlation between SABINS and CTS	0.719
Regression Coefficient on CTS Estimate in Predicting SABINS	0.810
N	*29,419*

Note: Income means and variances are calculated separately for SABINS boundaries and CTS boundaries using the method described in Section V Part C. Table reports separate correlation coefficients and regression coefficients comparing school-level income variables.

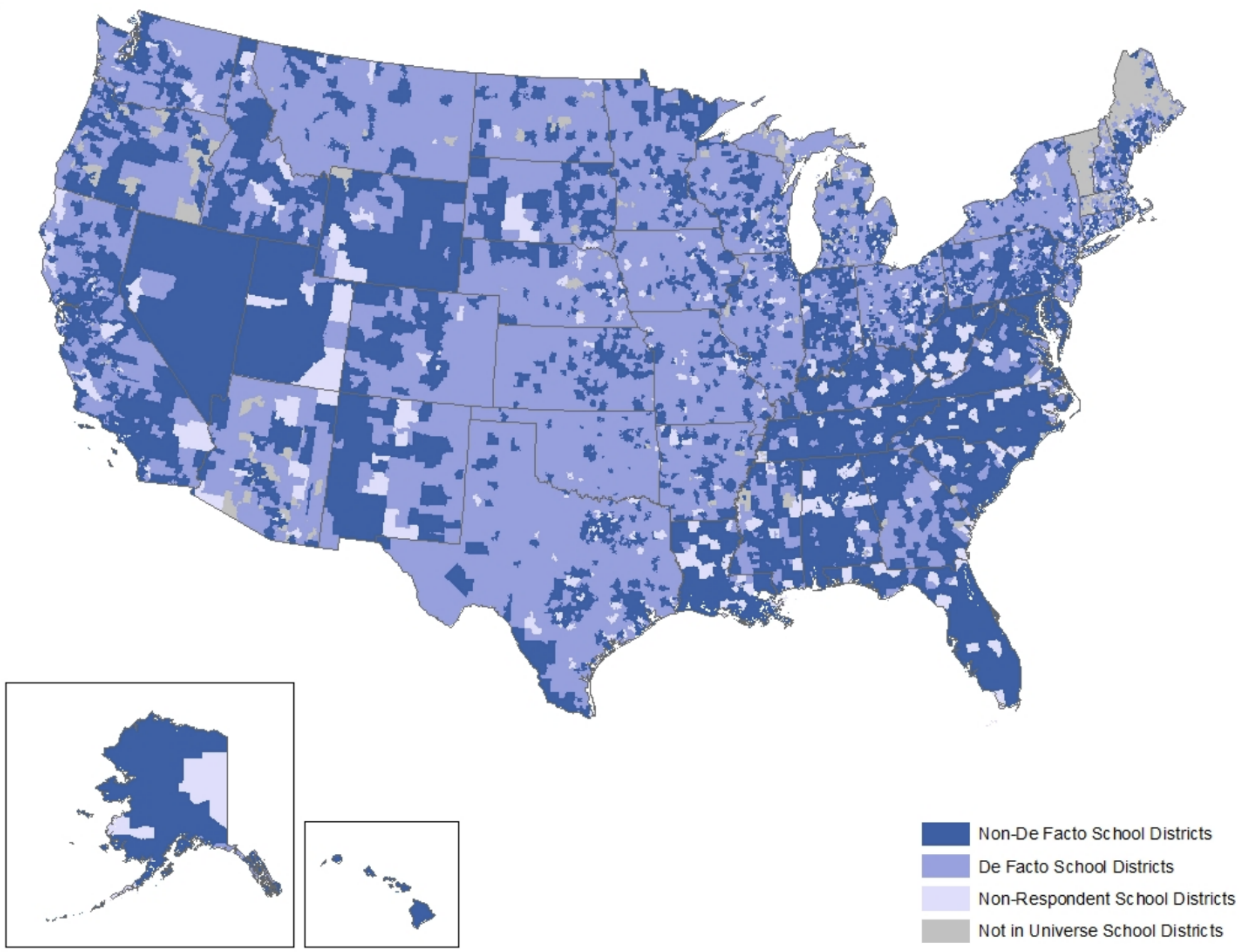
Source: The College of William and Mary and the Minnesota Population Center. (2011). School Attendance Boundary Information System (SABINS): Version 1.0. Minneapolis, MN: University of Minnesota.

Figure 1: SABINS National Coverage, 2009-10



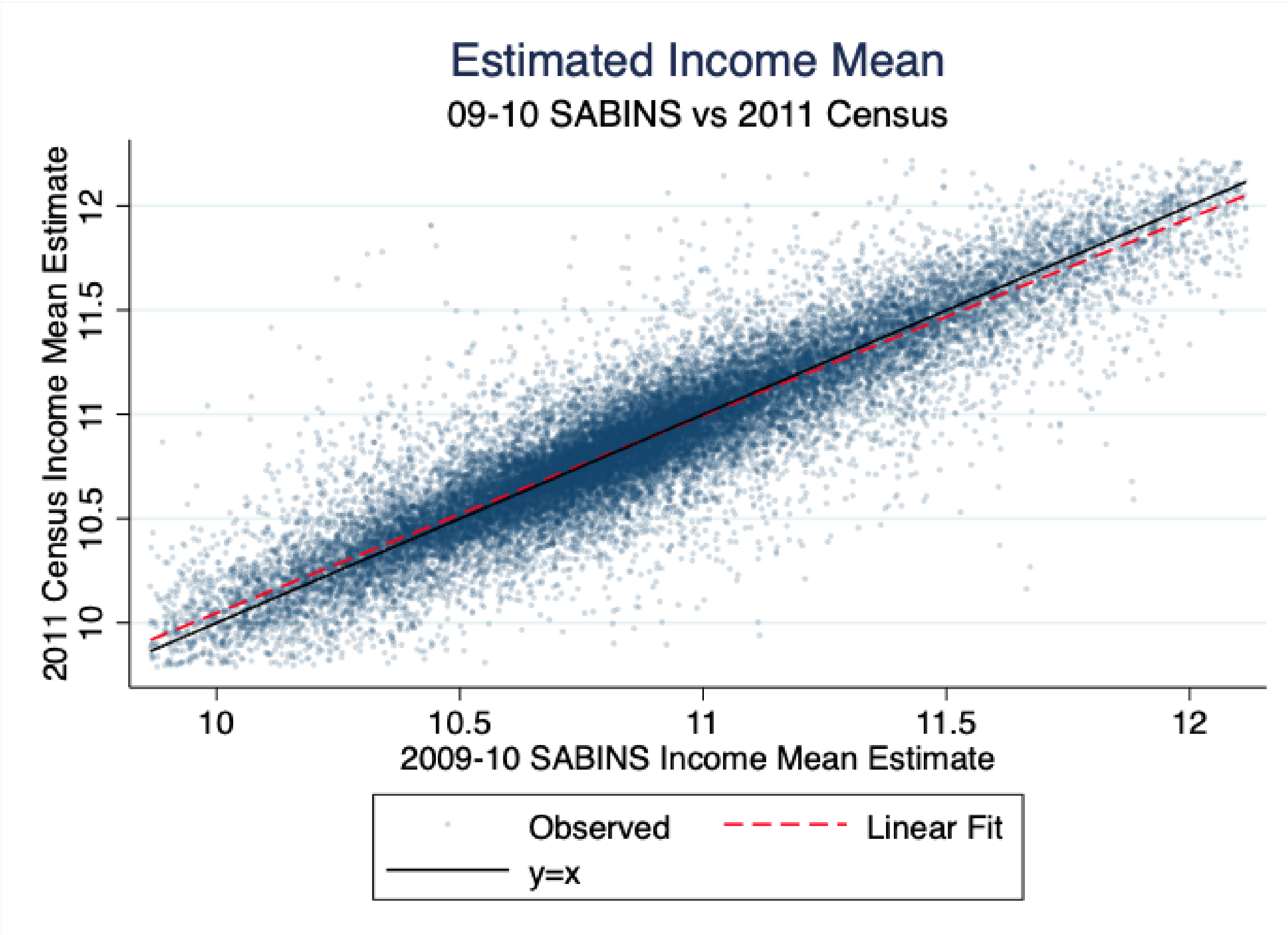
Source: Image from https://www.nhgis.org/sabins-data-availability.

Figure 2: SABS National Coverage, 2015-16



Source: Image from Geverdt, D., (2018).

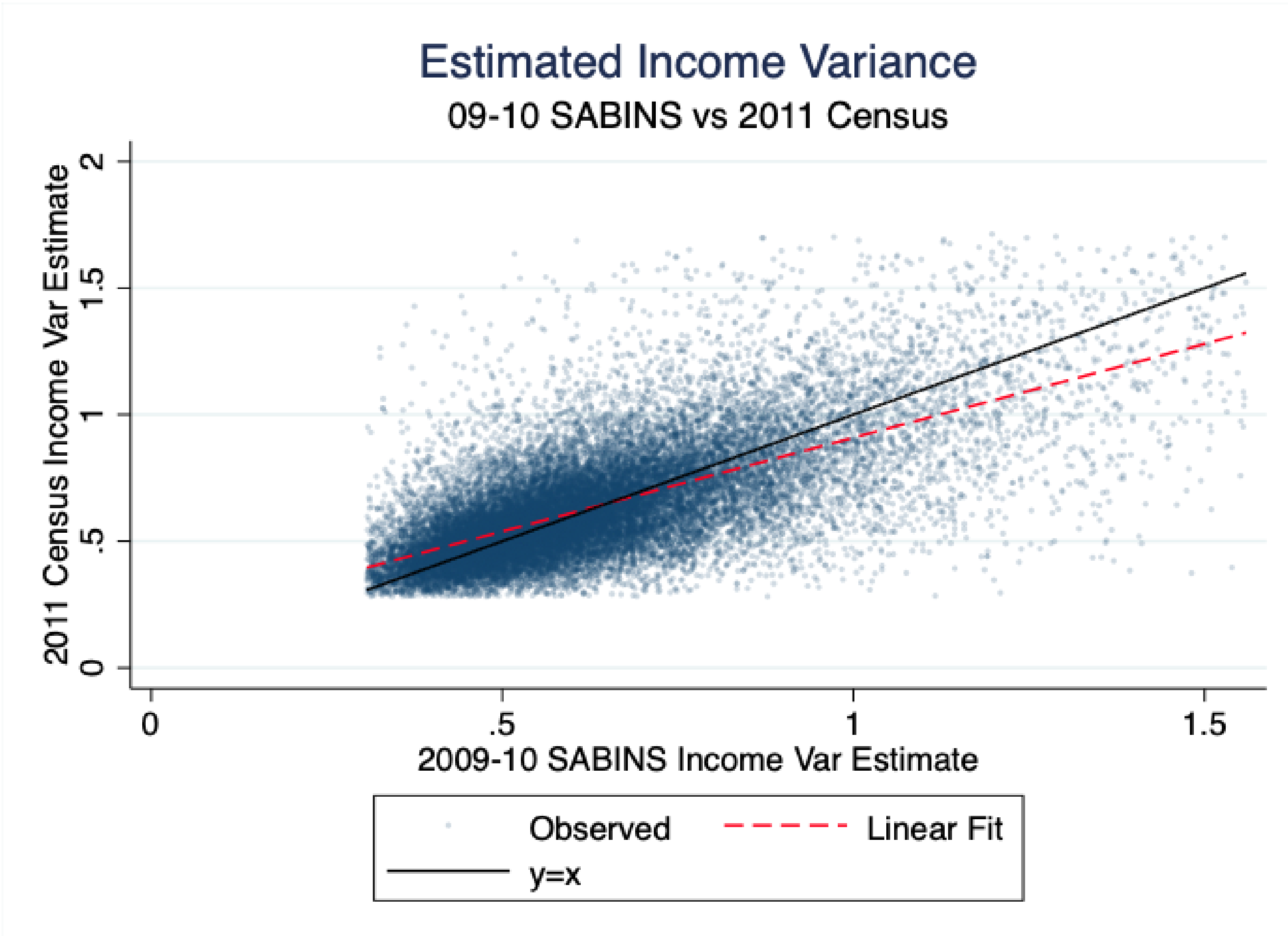
Figure 3: Estimated School-Level Income Means in SABINS 2009 and CTS Method using 2011 Census



Note: School-level income means are calculated separately for SABINS boundaries and CTS boundaries using the method described in Section V Part C. Each dot represents the estimates for a unique school. To constrain the impact of outliers, values in the first and ninety-ninth percentiles are trimmed.

Source: The College of William and Mary and the Minnesota Population Center. (2011). School Attendance Boundary Information System (SABINS): Version 1.0. Minneapolis, MN: University of Minnesota.

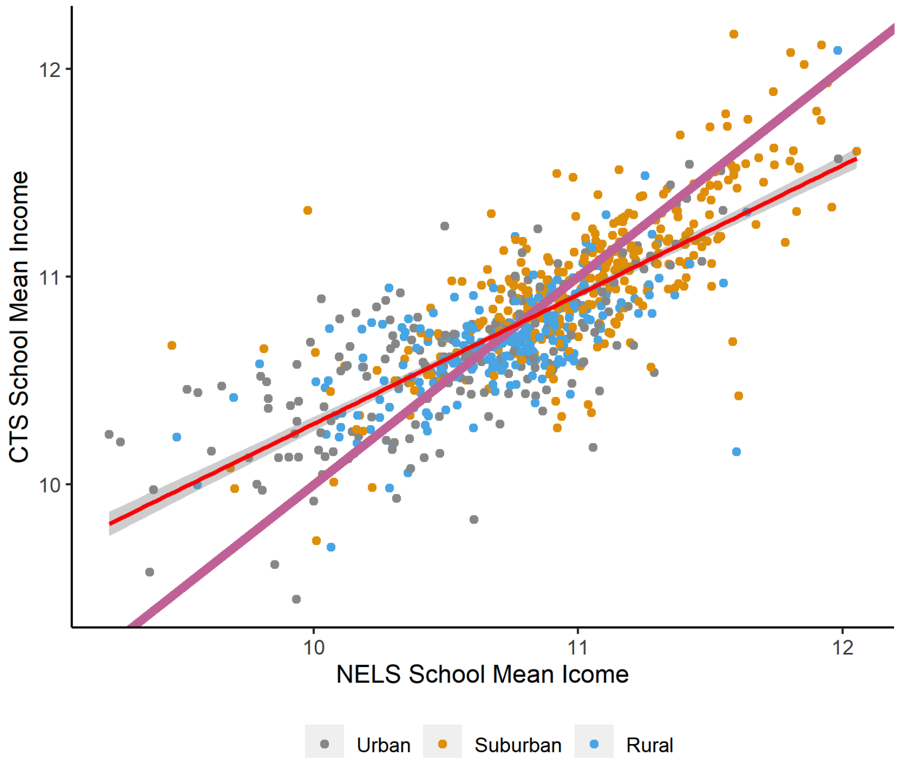
Figure 4: Estimated School-Level Income Variances in SABINS 2009 and CTS Method using 2011 Census



Note: School-level income variances are calculated separately for SABINS boundaries and CTS boundaries using the method described in Section V Part C. Each dot represents the estimates for a unique school. To constrain the impact of outliers, values in the first and ninety-ninth percentiles are trimmed.

Source: The College of William and Mary and the Minnesota Population Center. (2011). School Attendance Boundary Information System (SABINS): Version 1.0. Minneapolis, MN: University of Minnesota.

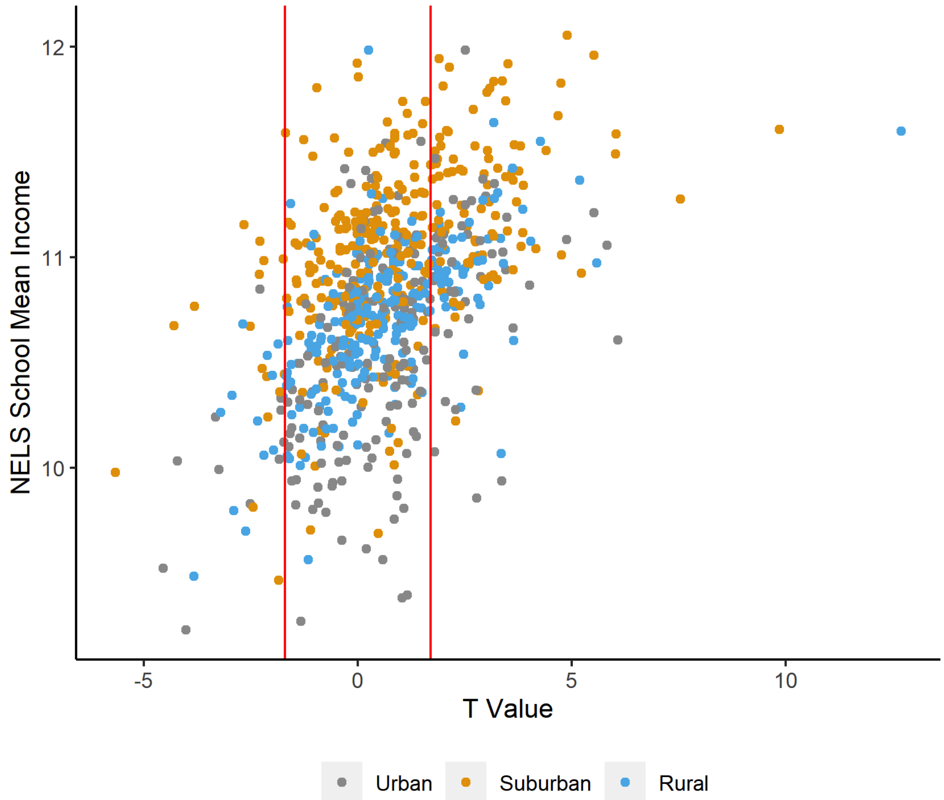
Figure 5: Estimated 1987 School-Level Income Means from NELS:88 and the CTS Method



Note: Each dot represents the school-level income mean estimate for a unique school. Means are calculated separately using the CTS boundaries and method described in Section V Part C (y-axis) and from calculating the simple mean of all reported income values in the NELS data (x-axis). In both scenarios, since the income data were reported in bin categories, we infer a school-level mean income based on the weighted proportion of students in each income bin (see text for further detail). The red line is the best linear fit; the bold pink line is y=x.

Source: United States Department of Education. National Center for Education Statistics. National Education Longitudinal Study, 1988.

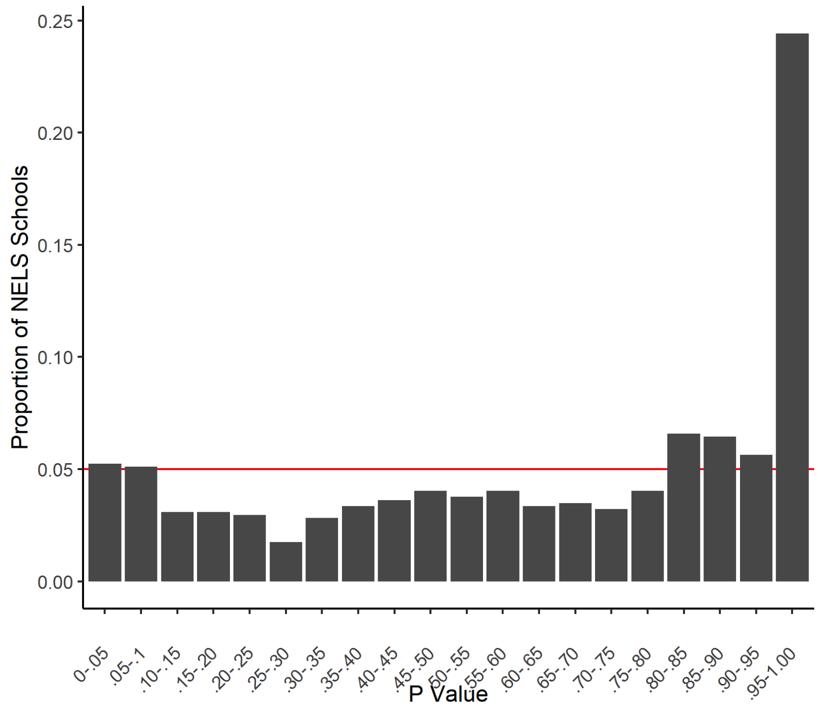
Figure 6: T-Statistics under CTS Null and Estimated 1987 School-Level Income Means from NELS:88



Note: Each dot represents the t-statistic for each school, assuming the CTS estimate of a school's income mean is in fact its true mean and testing the likelihood that we would observe a sample mean that is equal to or more extreme than the one we estimate using NELS. Red vertical lines indicate the critical values for the 95% confidence interval.

Source: United States Department of Education. National Center for Education Statistics. National Education Longitudinal Study, 1988.

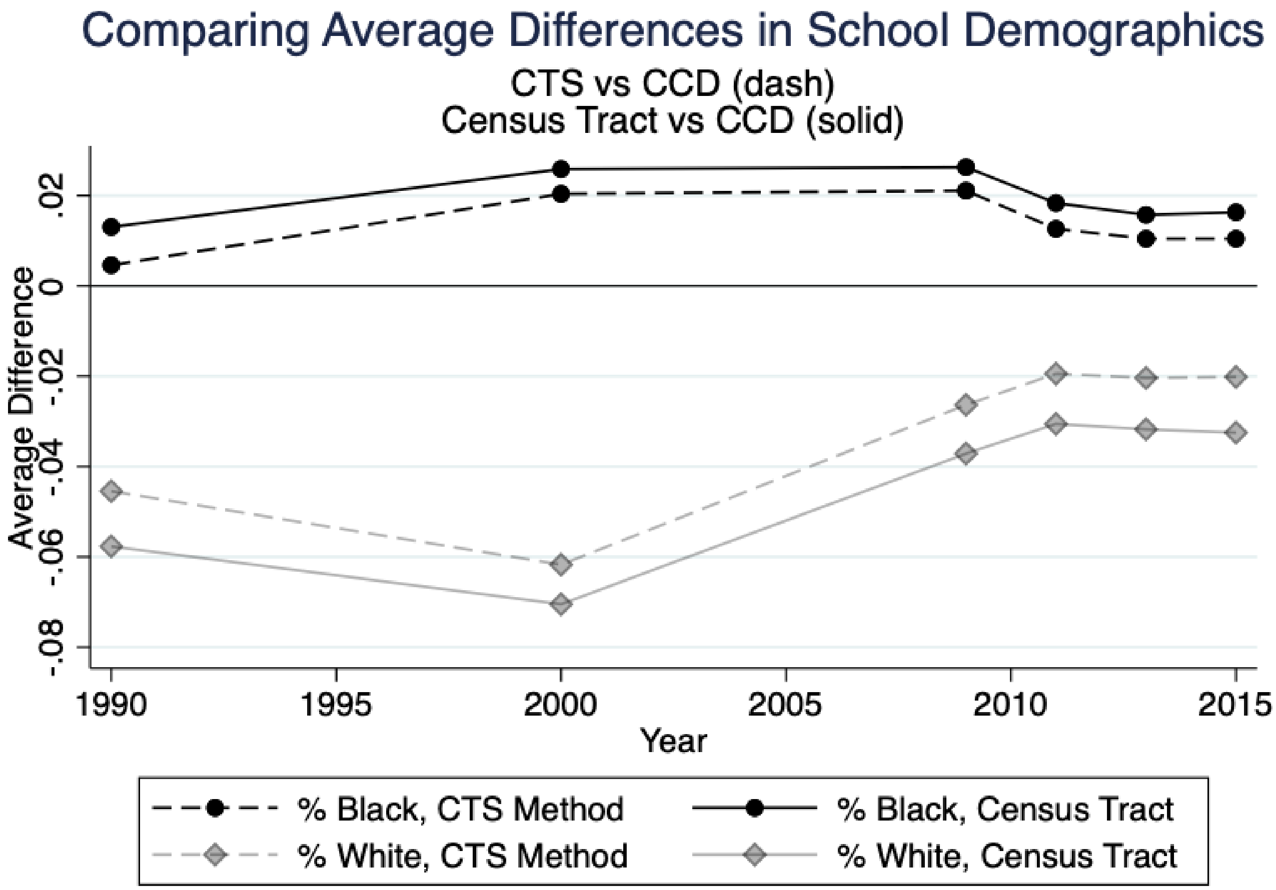
Figure 7: P-Values under CTS Null and Estimated 1987 School-Level Income Means from NELS:88



Note: Distribution of corresponding p-values for the t-tests presented in Figure 6. Under the null distribution, we would expect a uniform distribution with approximately 5% of schools falling into each of the 5 percentage point ranges of p-values as indicated by the red horizontal line. Bars above the red line at 5% represent an excess of p-values for that bin.

Source: United States Department of Education. National Center for Education Statistics. National Education Longitudinal Study, 1988.

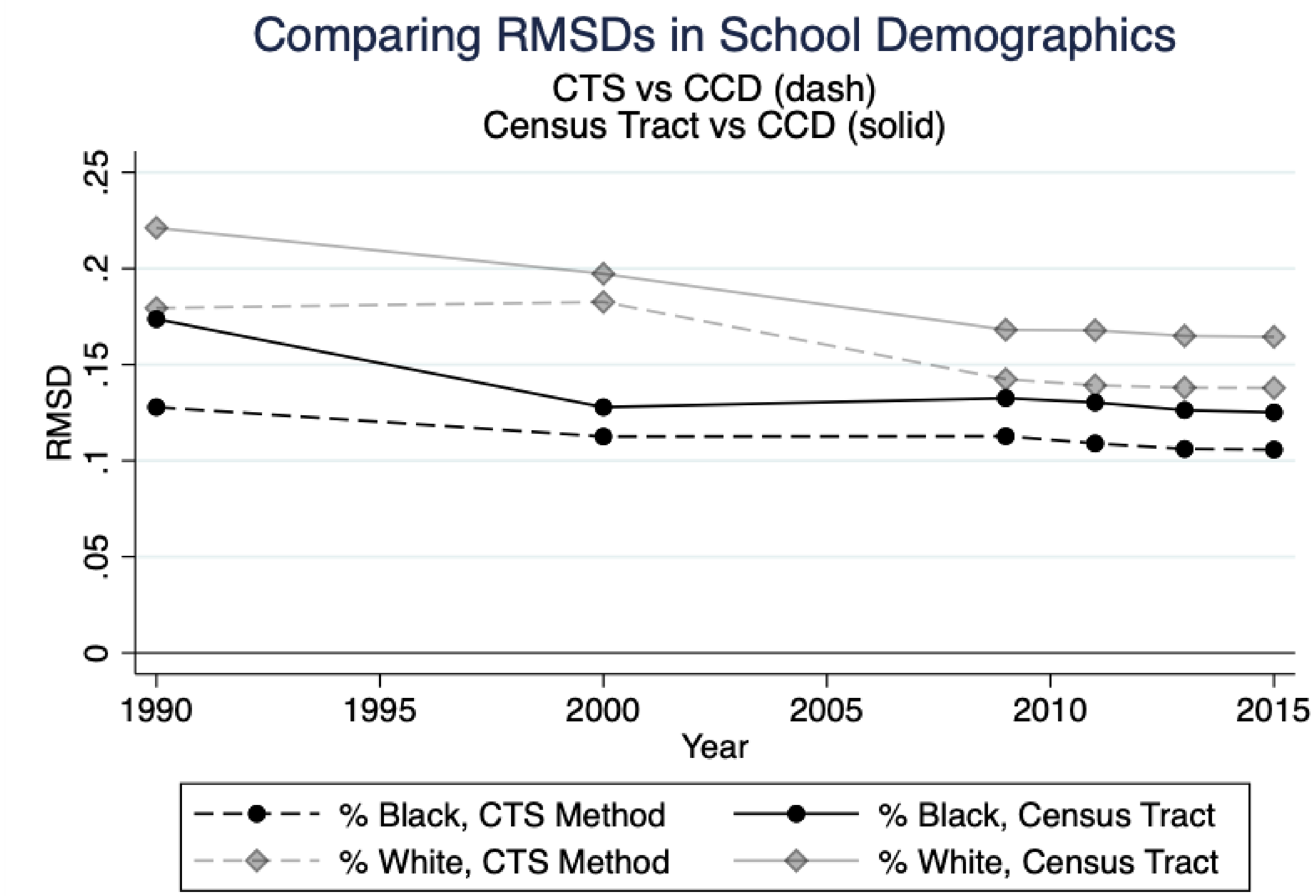
Figure 8: Average Difference in Estimates of CCD School Level Characteristics using CTS Method and Point-in-Place Approach



Note: For each school in each year, we estimate the percentage of students that are Black or White using our CTS method and the point-in-place census tract method. We then find the differences between these estimates and the true CCD-reported percentage and take the average of all of these school-level differences in each year. The closer the difference is to 0, the more accurate the estimates are, on average. The solid lines compare the estimate based on a point-in-place approach (i.e. the census tract in which a school is located) to the characteristic from the CCD. The dashed lines compare the CTS estimate to the characteristic from the CCD.

Source: U.S. Department of Education. National Center for Education Statistics. Common Core of Data: Public Elementary/Secondary School Universe Survey

Figure 9: Root-Mean-Square Deviations of Estimates of CCD School Level Characteristics using CTS Method and Point-in-Place Approach



Note: The RMSD can be interpreted as the average error that the estimate when compared to the actual school characteristic found in the CCD, where more weight is given to larger estimation errors. The closer the RMSD is to 0, the more accurate the estimates are, on average. The solid lines compare the estimate based on a point-in-place approach (i.e., the census tract in which a school is located) to the characteristic from the CCD. The dashed lines compare the CTS estimate to the characteristic from the CCD.

Source: U.S. Department of Education. National Center for Education Statistics. Common Core of Data: Public Elementary/Secondary School Universe Survey