Trends in the Academic Achievement Gap Between High and Low Social Class Children: The Case of Brazil

Martin Carnoy, Leonardo Rosa, and Alexandre Simões

Stanford University

Introduction

In every country of the world for which we have test scores, students from lower social class families average lower academic achievement in school than students with higher social class background (for example, OECD PISA, 2016). The reasons for the relationship between social class and achievement are complex, and the magnitude of the achievement gap between rich and poor appears to depend on several factors, including the degree of social inequality and social segregation in the society, how social class is measured, whether mediating variables such as race and location (rural/urban, for example), are accounted for in estimating the relationship, the level of schooling in which achievement is measured, and how schooling resources are distributed among students from different groups (for a review of this literature, see Sirin, 2005; Rothstein, 2005). Despite these complexities, there is considerable evidence that much of the achievement gap between higher and lower social class children is in place before children enter formal schooling (Jencks and Phillips, 1998; Rothstein, 2005), and that, at least in developed countries, differences in school resources may not contribute much to increasing or decreasing social class achievement gaps (Rothstein, 2005; Alexander et al., 2007).

Nevertheless, there is evidence based on human capital theory that increasing academic achievement for low-social class students can contribute to their higher educational attainment, higher productivity, and higher wages (Deming et al, 2016; Murnane, Willett and Levy, 1995). It is therefore only a small leap to claims that investing in higher student achievement is a powerful strategy to increase future average productivity and economic growth (Hanushek et al, 2013), and,

in turn, to broad acceptance of the notion that policies lowering achievement gaps between children from rich and poor families can reduce future social and economic inequality in adulthood (World Bank, 2018).

In this socio-political context, local and international policy analysts have focused on student achievement trajectories, as measured by state, national, and international tests (OECD PISA, 2016; Carnoy et al, 2015; Carnoy et al, 2017). This focus is mainly because of the implications student achievement may have for future economic development, but others have also focused on changes over time in achievement gaps between (especially in the U.S.) race/ethnic groups (Coleman, Campbell, and Hobson 1966; Jencks and Phillips 1998; Fryer and Levitt 2004, 2006; Rothstein 2005; Card and Rothstein 2007; Reardon and Galindo 2009; Reardon, Robinson-Cimpian, and Weathers 2014; Carnoy and Garcia, 2017; Musu-Gillette et al., 2016), and between social class groups (in the U.S., see Reardon, 2011; Reardon, Robinson-Cimpian, and Weathers 2014; Hanushek et al, 2019; internationally, see Carnoy and Rothstein, 2013; Carnoy et al, 2015). These achievement trajectories by race/ethnicity and social class serve as an indicator of how equitable the "opportunity system" is in a particular society.

The best known of this literature analyzes educational quality trends (as measured by test scores) in the U.S. in the past 50 years and changes in the achievement gap between students from low and high income families. In a widely discussed paper, Reardon (2011) estimated that there was a major increase in the test score gap since 1970 between students from the highest 10% earning families and everyone else, even as black-white and Latino-white differences tended to decline. He suggests that this increasing difference may be the result of increasing income inequality and the increasing gap between how much highest income parents and all other parents can invest in their children's academic-related activities. In contrast, Hanushek et al (2019) argue that the difference in school achievement is large between low and high income students, but that the difference has been stable over recent decades.

An advantage that this U.S. research has in tracking such achievement trajectories over time is that samples of students have been tested on comparable national student assessments since the 1970s. Students have also been identified by race/ethnicity, parent-education levels, and poverty levels, as measured by eligibility for free or reduced school lunch (Reardon, Robinson-Cimpian, and Weatherford, 2014). A major disadvantage of these data is that the two available indicators of social class (parent education and FRL eligibility) cannot differentiate adequately between families in the upper 10% or 20% of family resources and students from families with very low levels of academic and financial resources (Carnoy and Garcia, 2017). A second disadvantage is that the US data are collected by school-based sampling of students. Although it is possible to identify students within a given school and location in each test year, it is not possible to track schools or school districts over time. This makes it difficult to relate changes in school resources or socio-economic contexts to changes in student-level achievement gaps.

In this paper, we use a methodology similar to Reardon (2011) and to Hanushek et al (2019) to analyze academic achievement gaps over a ten year period in a large developing country, Brazil. Brazil has highly detailed data on students' achievement trajectories and their social class background over the past twenty years. After several decades of rapid enrollment expansion, essentially all Brazilian children 6 to 14 years-old are in school. The national debate on educational quality during this same period also triggered important changes in educational policy. The country created a national testing system called SAEB in the 1990s to monitor quality and, by 2005, this system covered almost all public schools. With these powerful measurement instruments, came a huge amount of public data. The Brazilian achievement data are available for a shorter period of time than in the U.S., but they have none of the disadvantages concerning social class measures or being able to identify schools and school districts over time.

Furthermore, Brazil is an interesting case for analyzing changing achievement gaps in primary and middle schooling because Brazil is one of the most unequal countries in the world.¹ Yet unlike the U.S., it has managed to reduce income inequality since the 1990s (Paes de Barros

¹ The income share of the top 10% of income earners is greater than 50% of national income (compared to 45% in the U.S.), and, according to World Bank estimates, the Gini index is one of the 10 most unequal in the world in 2015 (0.51 compared to 0.41 in the U.S.). However, Brazil's Gini index was above 0.60 in the late 1990s.

et al, 2007). If Reardon's speculations that rising income inequality in the U.S. contributes to growing achievement gaps, we may be able to observe the opposite trend in Brazil.

Surprisingly, to the best of our knowledge, there is no research that analyzes the trends in Brazilian education over time and compares the performance of different socioeconomic groups on these national tests. Neither is there in the existing Brazilian literature a coherent overall analysis of the proximate sources of student achievement inequality.

This paper is largely descriptive, and it addresses three important questions:

- 1. What are the trends in achievement for low and high socioeconomic pupils in Brazil in the last year of primary and the last year of middle school (5th and 9th grade)?
- 2. Are these trends explained by changes in observable economic and social contextual factors over time, such as social stratification among schools, the distribution of educational resources among schools, or the distribution of economic resources among municipalities?
- 3. Since states have major policy control over education, to what degree do these trends vary by state?

The methodology we use in our analysis of achievement gap trends, formally outlined in the next section, differs from most of the previous literature in two ways. First, we use panel data, not cross-section results, to analyze trends in achievement differences between students from low and high social class families. Most existing literature usually focuses solely on cross-section differences in academic performance. Important exceptions are some recent papers analyzing the U.S. case, such as Reardon (2011) and Hanushek (2019). However, the nature of the Brazilian data allow us to delve into more possible explanatory factors for trends in achievement gaps than possible with U.S. data.

Second, when considering the proximate sources of inequality in student achievement, we consider not only national differences, but also variation among municipalities and differences across and within states. Like the United States, Brazil is a federal country, and the federal

constitution vests legal responsibility for educational policy largely in the states, and somewhat in even more local units--in Brazil's case, municipalities. From a theoretical standpoint, therefore, it is important to compare educational variation among sub-federal political units. In this paper, we control for municipal variation and state fixed effects. We also focus on variation across states of achievement gap trends. The state level analysis allows us to observe patterns possibly related to varying state political and educational policy contexts.

I. Methodology

We combine six survey years of student background information and academic performance to investigate education inequality patterns in the Brazilian public school system over ten years (2007-2017). We measure inequality as the distance in performance between students at the top and bottom quartiles of the socioeconomic distribution and the top and bottom deciles of the socioeconomic distribution. Our methodology has three main components: 1) the selection of items that will be part of our socioeconomic measure; 2) the generation of an index indicating the socioeconomic status for each student in each of six test years, 2007-2017; and 3) the estimation of the test score gap between low and high socioeconomic status students and its trajectory over the period, 2007-2017.

The selection of items to the socioeconomic measure

We used two measures of socio-economic background (SES) for our estimates. The first is composed of a construct of an index of a student's reported articles in the home, and it is our preferred one. We used nine (9) home articles to construct the home possessions index--number of bedrooms; number of bathrooms; number of cars; number of computers; refrigerator; number of radios; number of television sets; DVD player; and washing machine. The index was constructed using Principal Components Analysis (PCA). The PCA weights the items to form the home possessions index.Our first measure of SES is therefore the home possessions index

This construct of students' social class does not include family's characteristics. To account for these characteristics, and to simultaneously check whether such family characteristics have a separate effect on students' test score trajectory we add them in our regression analysis as

covariates. The family characteristics included are whether the household includes a maid; whether mother is present in the family; whether father is present; whether there are housemates in the home; whether mother is literate; and whether father is literate.

In our second measure of socio-economic background, we combined the household items at home with the family characteristics in constructing an alternative PCA measure of SES. The list of these items and the methodology and results of the analysis are shown in detail in Appendix B. This second method of constructing the SES index follows the selection of items commonly used in other research (Hanushek, 2019 and PISA, 2017). There are two differences in the variables we could use in constructing SES indices and the variables used by Hanushek (2019) and PISA (2017). First, we used information about parental literacy instead of their level of education. This is because of the large percentage of missing values in the parental education variable, especially for 5th graders. Students in the 5th grade may not know their parents' last grade in school, but they are much more likely to know whether parents are able to read. The *Prova Brasil* survey does not include information on parent occupation, which seems to be a common feature in constructing SES indexes. However, information on parent occupation is correlated with items in the household and would only add marginally to our measure (Hanushek et al., 2019). We also did not have access to data on household income, but the quality of this information is questionable when reported by young students, and even by 9th graders.

We pursued these two different approaches of PCA weighted articles in the home and the PCA with the more complete list of items mainly because we were concerned about losing information. While the items in the second approach are more standard, in our PCA analysis of the more complete index, we noticed that the first component of the PCA did not capture the family variables (see Appendix B). This first component is traditionally used as the SES measure in other studies. However, family variables would be expressed by the second or third component of the PCA. For this reason, when including only the first dimension of the PCA with the standard approach, we were losing important variables that might explain the gap. Altogether the first approach was more transparent and clear, and has better properties using only one component. Importantly, as we mentioned above, although in our prefered approach we did not include the

family's characteristics as items in our SES construct, we added these characteristics in our regression analysis.

Our estimated achievement gaps over the six test years (2007-2017) shown graphically in our results section use the articles in home index controlling as our (preferred) measure of SES. Both the "articles in home" and the "standard PCA with extended items" measures of SES yield very similar results in achievement gaps over time and in our estimates of economic and social factors associated with the gaps. In our analysis, we mainly present results using the more straightforward articles in the home index and then, for comparison, show results using the longer version measure of SES.

Generating the socioeconomic index

We follow the typical methodology in the literature and use the Principal Component Analysis (PCA) to reduce the several variables into one construct of the composite measure. In short, the PCA transforms the original d-dimensional data onto a new k-dimensional subspace (typically $k \ll d$). The first principal component will have the largest possible variance, which is traditionally used as the index measure. We apply this method to the two sets of items discussed above from the student survey we selected, and, as is typical in the literature on PCA construction, we extract the first component from the PCA as a proxy of socioeconomic status. When we use the nine items of articles in the home, this gives us one estimate of the first component of PCA; the second estimate emerges from including a larger set of variables. Thus, in both cases, the PCA combines the socioeconomic items into a single, continuous indicator standardized over the population and comparable over time (more details in Appendix B).

A second step is to identify low and high SES students. To do so, we use the socioeconomic continuous index and draw the *national* distribution within each year-grade. We divided the distribution into quartiles, and defined the low socioeconomic students as those in the first quartile of the distribution (25th and below SES scores), and high socioeconomic students as those in the fourth quartile (75th and above SES scores). While we only show comparisons between low and high SES students, we include all students in our analysis. For our analysis that

compares differences between Brazilian states, we draw the SES distribution within each state and analyze the quartile gap *within* each state over the period, 2007-2017.

Because we aim to understand trends of the SES gap at different parts of the distribution, we also estimate the interdecile range, i.e., comparing student performance at the top and bottom 10% of the SES distribution. Having all these measures allow eight different perspectives to assess the Brazilian SES gap trajectory in the last decade, which is a combination of either the 75/25 gap or 90/10 gap in mathematics or Portuguese language for students in 5th or 9th grade. We do a similar inter-decile analysis within each state.

Explaining trends in achievement inequality

We use the following regression to investigate the trajectory in socioeconomic inequality in education:

$$Y_{ijmst} = \beta_1 SES_i + \left(\sum_{k=09}^{17} \Phi_t \times SES_i\right) + \beta_2 Fi + \beta_3 X_i + \beta_4 W_{jt} + \beta_5 M_{mt} + \delta_s + \theta_t + \varepsilon_{ijmst}$$
(1)

where Y represents the mathematics or Portuguese scores of student *i*, in school *j*, municipality *m*, state *s*, and year *t*. SES_i is a dummy that indicates the SES quartile in which the student's socio-economic background index falls. The interaction $\Phi_t \times SES_i$ is the interaction between each year of the cohort of the student *i* and quartile (or decile). We are particularly interested in the coefficients of these interactions since they will show us the patterns in the inequality trends. Notice that we cannot follow individual students over time. Therefore, we are analyzing trends for different cohorts of students.

In our regression, we add several groups of covariates to observe how controlling for each of these groups affects our original, unadjusted gap in test scores. Fi represents family characteristics covariate vector (listed above) and X_i represents the covariate vector that includes student age, gender and race. W_{jt} and M_{mt} are variables in the school level and municipal level, respectively. The municipality and school variables are listed in the Data section and in Appendix C. Both groups of variables vary over time since we included state fixed effects (δ_s) and time fixed effects (θ_i). Standard errors are clustered at the municipality level (which is roughly

equivalent to school districts in the U.S.). The omitted year is 2007 (the baseline) and the comparison group is either the bottom quartile or decile, depending on the model specification.

Data

We use data from the *Prova Brasil*, spanning 2007-2017. *Prova Brasil* is part of the Brazilian national assessment system, and its goal is to provide reliable measures on the quality of the public education system. The test is applied every two years. The data is publicly available from the National Institute for Educational Studies (INEP). We use the available data from 2007 through 2017. There are mainly two data sets that provide the information for our analysis: student survey data and test score data.

Student Survey Data

In *Prova Brasil*, INEP collects information about student characteristics, including their social and economic background. We use survey responses to build a yearly socioeconomic measure for all students in 5th and 9th grades. Because the surveys change from year to year, we applied item harmonization procedures to guarantee their comparability longitudinally . The harmonization process consisted of preserving only questions that appeared in all years and making sure they had the same number of possible answers. For instance, in 2007 the question "number of cars in the household" had four possible answers (None, 1, 2, 3 or more) while in 2013 there were five (None, 1, 2, 3, 4 or more). In this case, the last category in 2013 was grouped with the second to last category ("4 or more" became "3 or more") to match the structure of previous years. The other eight items went through similar harmonization processes. The pattern was mostly related to quantities, merging the top category to the one below. We removed some items that could be informative in identifying socioeconomic status because they did not appear in all surveys. A question about the number of vacuum cleaners in the household, for example, was available in the 2007 survey but not in following years.

Student Test Scores

Student scores are comparable over time, since INEP adopted the item response theory (IRT) in *Prova Brasil*. INEP has consistently assessed students in math and language

(Portuguese) in grades five and ninth. Since this study aims at understanding educational inequality growth across six test years of the data, we normalized scores based on the first year of our data set parameters. In this sense, all test scores were measured in terms of 2007 standard deviations, allowing year to year comparisons.

Sample Restrictions

In a typical year, approximately 5 million students, 2.5 million in each grade, participate in the *Prova Brasil* (see Appendix Table A-2 for a breakdown by year). Our analytical sample includes all students that took both tests between 2007 and 2017. Nonresponse rates also vary due to changes in the test sample design.² With this exclusion we have 25 million students in our dataset and 50 million test scores. Since our analysis also relies on student survey responses to construct the socioeconomic indicator, the sample was initially restricted to observations with non-missing values in the items used in the principal component analysis (PCA). For this reason, 34% to 50% of the observations were dropped, depending on the year, which reduced our sample, on average, from 5 to 3 million students per year. Note that the missing data rates in each of these sets of variables--student SES and student test scores--are not additive, since essentially all students without a recorded test score also had missing socioeconomic information.

To determine whether missing values in the socioeconomic items and the resulting greatly reduced size of the sample bias our estimates of test score inequality (75/25 and 90/10) over time, we used various imputation methods to estimate average scores with imputed values of student and family characteristics. We compared the results with various methods of imputing values to our estimates that did not impute values for missing observations . We preferred the modal value imputation of each variable. As we show, modal impution, random imputation, and the no imputation trajectories are only slightly different (see Appendix A).

 $^{^{2}}$ In 2007, INEP defined the eligibility criteria as urban schools with at least 20 students enrolled in each class of each grade. Two years later students in rural schools also participated and the minimum enrollment rule was reduced to 20 students in each grade (dropping the classroom requirement). This modification allowed for a larger number of schools to take the test, including those with a high nonresponse rate, which ultimately increased the percentage of missing information in our sample (see Appendix A).

Additional data

To check whether other factors at the school level or municipality could invalidate our analysis, we used a set of auxiliary data. At the municipal level, the variables include gross domestic product per capita, available from the Brazilian Institute of Geography and Statistics (IBGE)'s, and government expenditure on education, health, welfare programs, transportation, and public housing per capita, which is available from the Institute for Applied Economic Research (IPEA). In addition, we included the value transferred to poor families under the Bolsa Familia program, a cash transfer initiative promoted by the Brazilian government. This data can be found at the Brazilian Portal for Open Data. As with all the other variables mentioned, for the Bolsa Familia data, we used annual per capita amounts at the municipality level.

The school-level features were extracted from the school and teacher surveys that are part of the *Prova Brasil* dataset. For the same reason discussed previously, these datasets also went through a harmonization process. The main school level variables include average teachers' age, percentage of teachers with a graduate-level degree, the average percentage of curriculum content covered, and the percentage of teachers that have witnessed students going to class under the effect of drugs, but other variables were selected as well using the LASSO method as described in Appendix C. We also included a variable for school average student socioeconomic background as measured by the articles in the home index.

Results

Brazil, National Level Analysis

We estimated social class achievement gaps at the school and individual level, in math and language scores for 5th and 9th grades. As mentioned, our measure of test scores is standard deviations from a mean, where the mean of the 2007 test in either language or mathematics in each of the two grades. The gap is the difference in average test score measured in SD of students at the 75th and 25th percentile. We also provide a comparison in the trends for the gap in average test score of students at the 90th and 10th percentile. We estimate the gap separately for language and mathematics and separately for 5th and 9th grade.

Our results indicate that even as test scores rose in Brazil, the Brazilian SES-achievement gap for all students in both subjects is large and increased in the past decade, more in 5th than 9th grade. In this period, average achievement on both the 5th and the 9th grade *Prova Brasil* tests increased substantially, even for low-SES students. For example, the gain for lowest quartile SES students on the 5th grade language test was 0.7 SD, on the 5th grade math test, 0.5 SD, on the 9th grade language test, 0.6 SD, and on 9th grade math, 0.3 SD. However, the gains are not the same for children from different socioeconomic backgrounds. Wealthier students consistently made larger gains than poorer children, especially in 5th grade. In this section, we offer an overview of inequality in education at the country level, breaking down the analysis by grade and subject.

We first present graphs comparing the trajectories of "unadjusted" test scores for students in the bottom and the top of the SES distribution, using our preferred measure of SES, based on articles in the home. Second, we use OLS regression analysis, adding covariates, to estimate whether changes in various sets of covariates for students' family characteristics, individual characteristics, school resources, municipal resources and policies, and state level differences affect the trends in inequality. Third, we describe the heterogeneity of achievement gap across states.

Inequality trends in educational achievement - unconditional analysis

Figure 1a shows that, on average, the 25% poorest students scored 0.52 SD lower in mathematics than the 25% wealthiest students in 2007. That gap increased to 0.86 SD in 2013 and dropped slightly in later years. However, it remained larger than at the starting point in 2007. Thus, in 2017, the gap was 0.71 SD. Overall, the 75-25 math achievement gap for students in 5th grade increased 40% in ten years.

By definition, the 90-10 mathematics achievement gap was larger, but it followed a similar trajectory as the 75-25 gap (Figure 1b). In 2007, students in the bottom 10% SES group scored 0.66 SD lower than students in the top 10% SES group. This difference increased to 0.91 SD in 2017, which is equivalent to a 38% increase. The somewhat larger increase in the 90-10

gap suggests that the growth in achievement inequality tends to come from the more extreme ends of the socioeconomic distribution.

In terms of language scores, the 75-25 and 90-10 achievement gaps vary somewhat less year to year than in mathematics, but the size of the gaps is similar (Figure 1c and 1d). In 2007, performance was 0.48 SD lower for the 25% poorest students from those in the 75 percentile, and 0.61 SD lower for the bottom 10% compared to the top 10% SES. In 2017, the gaps increased to 0.71 SD and 0.89 SD, respectively.

In Figure 2 (achievement gaps 75-25 and 90-10, 9th grade), we estimate the mathematics test score gap for students in 9th grade. At the beginning of the period covered in the analysis, math achievement inequality for these middle school students looked fairly similar to that of 5th graders. However, the gaps for 9th graders of high and low social class increased less than for high and low social class 5th graders. The average difference in math performance between those at the 75th and 25th percentile of the socioeconomic distribution was 0.55 SD in 2007 (Figure 2a). Ten years later, this number was approximately 7% larger, about 0.59 SD in 2017. The 90-10 comparison (Figure 2b) shows a similar trajectory as the 75th to 25th percentile comparison. The achievement gap for the top versus bottom SES decile increased from 0.69 SD to 0.72 SD, a 4% increase during the 2007-2017 decade. The results for 9th grade language score gaps indicate an insignificant increase in the 75-25 achievement gap (Figure 2c) from 0.44 SD in 2007 to 0.45 SD in 2017 and a decrease from 0.55 SD to 0.54 SD in the 90-10 gap (Figure 2d).

Thus, our results indicate that mathematics and language achievement differences between students in the highest and lowest quartiles and deciles of the SES distribution appear to have widened over the last decade for 5th graders, and increased slightly or declined slightly for 9th graders . The gap has tended to widen more at the extreme ends of the distribution, as evidenced by the greater widening of the gap between the highest and lowest 10% of social class students than between the top and bottom quartiles of students.

The jump in the 5th grade 75-25 and 90-10 gaps in 2013 raises some issues about the 2013 results for that grade. We checked gains in mean scores over time for 5th and 9th grade students to confirm our regression results. These are shown in Table 1. The table shows that lower SES 5th

grade students made no gains in test scores in 2013, whereas higher SES students continued to make gains as in previous years. In 2015, lower SES students "bounced back" with very high gains. This suggests that something may have been amiss with the lower SES scores in 2013. Delving deeper into this problem requires further research.

How Sensitive are the Gaps to Imputing Missing Values and to Different Methods of Imputation?

In Appendix A we explain our various corrections for missing values and show a comparison of the average test score gap for the highest and lowest SES groups among 5th graders using different imputation methods. We end up using the mode per school as the chosen imputation method for missing values. In this mode per school case, there is no significant loss of data due to the missing values in the socioeconomic variables. In all years around 99% of all students with both language and mathematics scores were preserved in the analytical sample. The estimated achievement gaps estimated by modal and random imputations are only somewhat different from the estimates with no imputation, but by including a much higher percentage of respondents through imputing the missing responses to various questions, we are more confident that our estimates are representative of the entire student population.

How do inequality trends in educational achievement change when adding context factors in schools, municipalities, and states?

In the next level of analysis, we use the composite measure with the articles in the home proxy to estimate the SES gap in Tables 2-5. In these tables we also explore how students' family characteristics, student characteristics (race/gender/age), and school, municipal, and state factors, including resource differences in schools and municipalities, might "explain" the changes over time in the gap. We do the same analysis but in less detail in Tables 6 and 7, where we also include the gap estimates using the full PCA SES measure for comparison purposes. The estimates provide considerable information on how these contextual factors may have contributed to the social class achievement gaps over this period of time.

We focus on Tables 2-5. These present a detailed set of estimates of the highest to lowest quartiles SES achievement gap using home possessions as our measure of SES. The coefficient of

the gap represents the coefficient of the dummy denoting students whose score fell into the top quartile of test scores in a given year compared to students whose score fell into the bottom quartile in that same year. To make the table easier to read, we omitted the coefficients of the second and third quartile dummies compared to the first quartile, but they are included in the regression as well as all quartile-year interactions. The key results are the Pct4 coefficients--which represent the 75-25 achievement gap in the bse year, 2007--and the coefficients of the various years interacted with PCT4--these represent the difference in the 75-25 achievement gap between the given year, say 2013, and the base year gap.

There are several important insights we gain from these tables. The first relates to the impact on the level of the gap in the base year from adding covariates for family characteristics as controls to our articles in the home SES measure. That tells us how robust our measure of SES is using just the weighted articles in the home as our SES measure (coefficient of Pct4 in column 2 compared to column 1 in each table). Making this comparison for language and math scores in the 5th grade (Tables 2 & 3, column 2 Pct4 versus column 1 Pct4) suggests that for 5th graders, controlling for family characteristics does not change the gap significantly--hence the articles in the home index provides a robust measure of SES for these purposes. However, this is not the case for grade 9 (Tables 4 & 5, column 2 Pct4 versus column 1 Pct4). Controlling for family variables significantly reduces the gap compared to using only articles in the home to measure SES. This suggests that for 9th graders, family characteristics are picking up elements of social class as it relates to achievement that the articles in the home index does not.

The second insight is that controlling for individual characteristics, such as race, gender, and age, significantly reduces the estimated achievement gap in both 5th and 9th grade. This means , for example, that there is a higher percentage of lower scoring race groups in the bottom quarter of SES, and that this non-social class factor explains part of the gap in test scores between students in the top and bottom SES quartiles. There are many reasons that Indigenous Brazilian students or African-Brazilian students may score lower on tests, but if these reasons are not social class related, we should adjust for race in tracking the SES achievement gap.

The third insight is that school factors--the average SES of students in the school and school resources--especially the average social class of the school, the variation of which is an approximate indicator for school social class segregation--seems to "explain" a large fraction of the inter-SES quartile achievement gap in Brazil (compare Tables 2-5, columns 4 and 5 with both columns 1 and 3). This is true for both 5th and 9th graders, and more so for 5th than 9th graders. This means that much of the achievement gap can be accounted for by between school differences, and it suggests that if a student is from a low-SES home and attends a school with low average SES, he or she will perform relatively more poorly on the *Prova Brasil* test than it he or she had attended a higher average SES school. Why that is the case is a complex issue (see Carnoy and Garcia, 2015), yet it appears that social class segregation in Brazil had a negative effect on achievement equality.

A fourth insight is that relative to the effect that school factors have on achievement inequality, the variation in municipal resources and policies not captured in the school variables appear to have had little impact on achievement inequality in both 5th and 9th grades--even less in 9th than in 5th (Tables 2-5, column 7 compared to column 6).

We can also draw some tentative conclusions about the effects of these different covariates on achievement inequality over time. As an example the results for mathematics are summarized in the graphs in Figures 4 and 5. The results for language in each grade are very similar. First, SES achievement inequality increased substantially in 2007-2017 among 5th graders but not 9th graders, as already noted. Yet, student personal characteristics, such as race, gender, and age, are less of an "explanatory" factor of SES achievement inequality in 2017 than in 2007 in both 5th and 9th grade. This may reflect relative gains, for example, by lower scoring race/ethnic groups.

Similarly, school average social class and context/resources become less important over this period in "explaining" the 75-25 achievement gap, especially among 5th grade students and especially after 2009. The decrease in school factors contribution to SES achievement inequality is much smaller for 9th graders and occurs later, after 2013. These results suggest that SES

differences and resource differences between schools became somewhat more equal in 2007-2017, particularly in fifth grade.

On the other hand, controlling for municipal economic differences in addition to school level differences had little impact on the increases in the gap in either grade. This suggests that municipal economic differences and municipal policies (spending on schools and social services, including Bolsa Familia) were relatively neutral over time in this period vis-a -vis SES achievement gaps, once changes in school context were accounted for. Finally, state fixed effects, which measure differences in a whole range of time-invariant factors among states, also explain little about changes in the trajectories of achievement inequality over time when we account for individual and school factors..

Apparently, disparities among states relevant to SES achievement gaps became greater especially after 2013. With all these controls, the 75/25 achievement gap in 2017 had increased by about 0.2 SDs over 2007 in both language and math, about the same when we include family characteristics in our measure of SES and control for no other variables. In our estimates for 9th grade language and math scores, the influence of school level and municipal level factors on the SES achievement gaps in language and math behaves similarly to their influence on 5th grade gaps (Tables 4 and 5). The main difference between the analysis for 5th and 9th grade SES gaps is that the gaps for 9th graders increased much less than for 5th graders, and almost the entire increase came in 2013. In language, once we controlled for school, municipal, and state factors, the gap in 2017 was on 0.04 SDs higher than in 2007, and in math, only 0.09 SDs higher than in 2007.

How related are the estimates of inequality trends and context factors to different measures of social class?

As noted, we developed two measures of student social class—an index of 9 articles in the home and a PCA index which uses more variables.. Tables 2-5 show the estimates of the changing language and mathematics gaps between the 25th and 75th percentile of SES for 5th and 9th graders over the period 2007-2017, when we use an articles-in- the-home measure of SES. For comparison with the results using the PCA analysis (below), we refer to the regressions in Tables

2-5 with no controls (column 1). For example, that estimate in Table 2 shows an increase in the gap between 2007 and 2013 of 0.28 SDs for 5th grade language scores and a decline in 2013-2017. For 5th grade mathematics (Table 3), the gap increases by 0.33 SDs in 2007-2013, and also declines in 2013-2017.

For 9th grade students (Table 4 and 5), the estimated increases and decreases in the gap are small throughout the years 2007-2017. The gap essentially did not change in 2007-2017 (column 1).

When we compare the results in Tables 2-5 with the estimates of the gap over time, using the extended PCA composite measure for SES (Table 6 and 7), we get very similar results. We compare the articles in the home regressions in Table 1-4 with the Full PCA composite measure of SES, and we similarly compare the regression estimates with all covariates (Tables 2-5, column 8 with the corresponding regression estimates using full PCA. This comparison shows, generally, that the differences between estimates using the two somewhat different constructs of SES are small. Adjusting the estimates for all covariates does have a bigger effect on the initial achievement gap in 2007 for 5th graders when the full PCA measure of SES is used. Other very small differences also emerge.

These small differences in the estimates of changes in the varying gap suggest that whether we use the extended measure of SES or only articles in the home the estimated pattern of the changing test score inequality gap in Brazil during this period is essentially the same.

State Differences

<u>Fifth Grade.</u> Analyzing the educational achievement gap at the country level hides variation among states. Figure 3 shows changes in state average learning gains for 5th graders from 2007 to 2017 against the increase in social class achievement inequality over the same period. All results are estimates for states using the model specification from column 1 in Tables 2-5. Student mathematics scores in 5th grade improved by about 0.7 SD between 2007 and 2017, and, in language, about 0.9 SD. Improvements varied from state to state.

The best case scenario--if we believe that high gains in test scores accompanied by decreases in test score inequality between low and high SES students is socially optimal--is in states that improved performance while reducing disparities between low and high SES students. Students in most states increased their average achievement level, but all also increased the level of inequality between students from high and low socioeconomic status. Sao Paulo was the only state that improved 5th grade scores without widening the socioeconomic gap, but even in this state, the gap did not fall significantly (the estimated variation is not statistically significant from zero). Nevertheless, this state was an isolated case. In contrast, most north and northeastern states widened the learning gap more than the national average.

These results suggest that improvements in learning outcomes (as proxied by the *Prova Brasil* test scores) were more concentrated in already economically advantaged students rather than the disadvantaged. One example is the state of Ceara. It stands out for improving student performance on the SAEB in this period more than other regions, but, as was the case more generally in Brazil, it did so at the expense of an increase in student performance inequality--the social class achievement gap increased almost as much as the national average. Another example is the state of Amazonas. Of all states, regardless of subject (math or language) or SES group comparison (inter-quartile or inter-decile), the greatest increase in the social class achievement gap was in Amazonas. Even though student test performance in Amazonas increased slightly more than the national average, the poorest 25% students in that state fell almost 0.5 SDs behind the richest 25% students during the ten year period. This increase in the 75-25 gap was approximately 3 times that experienced by students in the country as a whole during the same time frame.

Most northeastern states witnessed a similar trajectory in terms of inequality as Amazonas, but without the same level of learning gains. For instance, students in Maranhao increased their average math score and the interquartile range by the same amount (0.3 SD). In this sense, improvement in student performance in all these states was weighted toward wealthier students--precisely the opposite result from what policy makers would be aiming for were they

trying to close disparities in the outcomes of primary education between the haves and the have-nots.

<u>Ninth Grade.</u> Student math and language scores on the national test for 9th graders increased less than for fifth graders in 2007-2017. Average 9th grade math scores increase 0.3 SD and language score 0.6 SD. As noted, the social class achievement gap also increased much less among 9th graders. However, although overall improvement in scores was not as great for 9th graders, unlike in the case of 5th graders, we estimate that in some states, the socioeconomic achievement gap between lower SES and higher SES 9th graders declined. For example, low-SES students in the state of Paraiba increased their math and language scores approximately 0.1 SD more than high-SES students in 2007-2017, and low SES 9th grade students in the state of Rio de Janeiro also tended to improve relative to high SES student while, overall, student test scores increased during this period.

Other states also diminished learning disparities, but not as much or as consistently across subjects as Paraiba. For instance, low SES students in Sao Paulo state made 0.05 SD greater math gains than students at the top of the SES distribution, but not significantly greater gains in language scores. The SES pattern of student test score gains in the state of Rio de Janeiro is similar to Paraiba's, but the inter-decile and quartile gaps declined by only half (0.05 SD) of those in Paraiba. As noted for 5th graders, the state where the SES test score gaps increased most was Amazonas; in 9th grade, the greatest increase in achievement gap occurred in Acre and Mato Grosso. While Acre improved math and language learning above the national average, those gains were very uneven among disadvantaged and advantaged students. Students in the top 25% of the SES distribution in that state increased their SAEB performance 0.2 SD more than students in the bottom 25% of the SES distribution (in both math and language). Similar to the pattern for 5th graders in Maranhao, the achievement gap between low- and high-SES 9th graders in Maranhao increased even as their average test scores made relatively small gains. Others states such as Para, Amapa, and Mato Grosso show similar patterns for 9th graders as in Maranhao. All four states increased scores below the national average while still increasing the social class achievement gap substantially.

Charting state differences in test score gains versus SES achievement gap increases controlling for school and municipal factor variation changes the states' positions on the graph somewhat but not meaningfully. Fundamentally the inferences we draw about how some states are located in the quadrants and our overall conclusions regarding the tendency toward 5th and 9th grade test score improvements with increasing SES achievement inequality does not change by controlling for additional factors. These results are available on request.

Discussion and Conclusions

Our results suggest that social class inequality in public school student achievement on the national test increased significantly among Brazilian 5th graders in the period 2007-2017, and that all of this increase was before 2013--indeed, most between 2011 and 2013. Since 2013, the gap has diminished. This is the case for both language and mathematics scores even when we account for changing 5th grade students' family characteristics and personal characteristics during this period. Although we were able to infer that school, municipal, and state factors were related to this increase in the social class achievement gap in particular ways, including these various covariates drastically reduces the size of the achievement gap, but does not change the main trends appreciably.

On the other hand, social class achievement inequality among 9th graders barely increased in 2007-2017, and that small increase appears to have occurred mainly in earlier years, then after a decline, increased again in 2015-2017.

The relationship of school, municipal, and state factors to the 9th grade gap is similar to that in 5th grade, namely that personal characteristics such as race are related to the achievement gap and this is also the case for inter-school student social class and school resource inequality. Both personal and school characteristics had a decreasing explanatory effect on the achievement gap over the period, suggesting that in the past ten years, either race/gender/age and school social class and school resource differences all tended to decrease or that their impact on achievement decreased.

We also were able to infer that municipal economic differences and that spending on schooling and other social services among municipalities in this period had relatively little impact on the achievement gap once school factors were accounted for, and that state factors became more somewhat more unequal, especially after 2013, which contributed to slightly greater social class inequality in test scores. It appears, then, that public school segregation and resource distribution or other policies (such as moving the school starting age from 7 to 6 years old)³ and municipal public spending policies may have contributed to reducing the social class achievement gaps in both 5th and 9th grade during this ten-year period, and that other factors, which we have not been able to measure, such as within-school differentiation may have come to play a larger role in increasing social class achievement differences, especially in primary school.

Nevertheless, the question remains why the fifth grade social class achievement gap increased by 0.2 SDs and the 9th grade gap did not. We can infer from our regression analysis, which controls for variation of resources among schools, among municipalities, and includes state fixed effects, that the 5th grade social class achievement gap was influenced considerably by between-school inequality during this period, much more at the beginning of the period than at the end. This earlier period was one of relatively high economic growth and declining income inequality; thus, the question is whether in this period schools all over Brazil followed policies in these years that may have increased differences in achievement among fifth graders of higher and lower social class within the schools, or perhaps that an external policy, such as pre-school provision in the years before 2007, or the availability of after-school or vacation-time activities during this period may have been applied in a way that increased differences in student achievement among higher and lower social class students in the same school.

These "mysteries" are important to unravel, since the magnitude of the increases in achievement inequality among fifth graders are large. they therefore may have important implications for the relative opportunities available to lower social class students. On the other hand, any such research must also explain why the social class achievement gap did not increase

³ See Rosa et al, 2019.

among ninth grade students. For example, it is possible that dropouts in middle school occur mainly among the lowest achieving students in the lowest quartile or decile of student social class. Thus, even as more of these students entered middle school during this period, they dropped out by 9th grade, eliminating an important source of increasing test score inequality. In 2015-2017, this trend reversed, perhaps the result of declining dropouts in middle school. If this is the case, we should observe a continuing increase in 9th grade achievement inequality in the 2019 SAEB.

References

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2007). Lasting consequences of the summer learning gap. *American sociological review*, 72(2), 167-180.

Card, D., & Rothstein, J. (2007). Racial segregation and the black–white test score gap. Journal of Public Economics, 91(11-12), 2158-2184.

Carnoy, M. & Rothstein, R. (2013). What do international tests really show about US student performance? Economic Policy Institute.

Carnoy, M., Garcia, E., & Khavenson, T. (2015). *Bringing it back home*. Washington, DC: Economic Policy Institute, (EPI Briefing Paper, No. 410).

Carnoy, M., Marotta, L., Louzano, P., Khavenson, T., Guimarães, F. R. F., & Carnauba, F. (2017). Intra-national comparative education: What state differences in student achievement can teach us about improving education—the case of Brazil. *Comparative Education Review*, *61*(4), 726-759.

Carnoy, M., & Garcia, E. (2017). Five Key Trends in US Student Performance: Progress by Blacks and Hispanics, the Takeoff of Asians, the Stall of Non-English Speakers, the Persistence of Socioeconomic Gaps, and the Damaging Effect of Highly Segregated Schools. Economic Policy Institute.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., & Weinfeld, F. (1966). Equality of educational opportunity study. Washington, DC: United States Department of Health, Education, and Welfare.

Cowen, C.D. et al (2012). *Improving the measurement of socioeconomic status for the NAEP: A theoretical foundation*. Washington, D.C.: National Center for Educational Statistics. <u>https://files.eric.ed.gov/fulltext/ED542101.pdf</u>.

Deming, D. J., Yuchtman, N., Abulafi, A., Goldin, C., & Katz, L. F. (2016). The value of postsecondary credentials in the labor market: An experimental study. American Economic Review, 106(3), 778-806.

Fryer Jr, R. G., & Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. Review of economics and statistics, 86(2), 447-464.

Fryer Jr, R. G., & Levitt, S. D. (2006). The black-white test score gap through third grade. American law and economics review, 8(2), 249-281.

Hanushek, E. A., Peterson, P. E., & Woessmann, L. (2013). *Endangering prosperity: A global view of the American school*. Brookings Institution Press.

Hanushek, E. A., Peterson, P. E., Talpey, L. M., & Woessmann, L. (2019). *The unwavering ses achievement gap: Trends in us student performance* (No. w25648). National Bureau of Economic Research.

Jencks, C., & Phillips, M. (1998). The black-white test score gap. Washington, DC: The Brookings Institution.

Murnane, R. J., Willett, J.B., & Levy, F. (1995). The growing importance of cognitive skills in wage determination." *Review of Economics and Statistics*, 77(2), 251-266.

Musu-Gillette, L., Robinson, J., McFarland, J., KewalRamani, A., Zhang, A., & Wilkinson-Flicker, S. (2016). Status and Trends in the Education of Racial and Ethnic Groups 2016. NCES 2016-007. National Center for Education Statistics.

OECD, PISA. (2016). Results (Volume I): Excellence and equity in education. Paris: OECD.

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. *Whither opportunity*, New York: Russell Sage Foundation. I(1), 91-116.

Reardon, S. F., & Galindo, C. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. American Educational Research Journal, 46(3), 853-891.

Reardon, S. F., Robinson-Cimpian, J. P., & Weathers, E. S. (2014). Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps, in Ladd, H. & Goertz, M. (eds.). *Handbook of Research in Education Finance and Policy*. Lawrence Erlbaum.

Rosa, L., Martins, M., & Carnoy, M. (2019). Achievement gains from reconfiguring early schooling: The case of Brazil's primary education reform. *Economics of Education Review*, 68, 1-12.

Rothstein, R. (2005). Class and schools. New York: Teachers College Press and the The Economic Policy Institute.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. Review of educational research, 75(3), 417-453.

World Bank (2018). *World development report: Learning to realize education's promise*. Washington DC: World Bank.

FIGURES

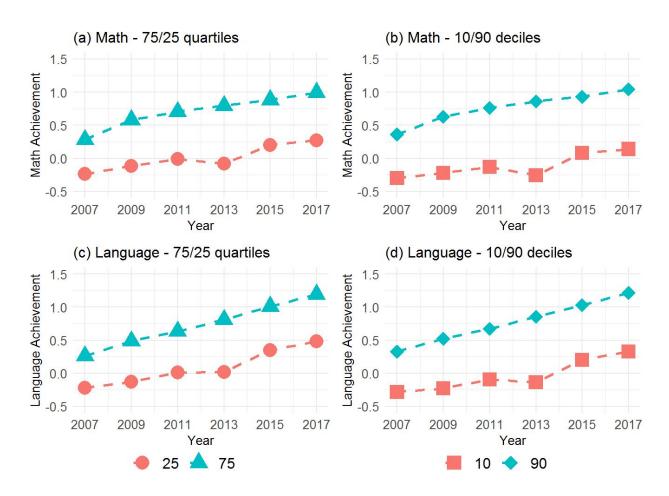


Figure 1. Brazil: 5th grade Achievement Gap Trajectories, 2007-2017

Notes: This panel reports unconditional achievement gaps for language and math test scores between students in fifth grade considered high and low socioeconomic status (SES). Students considered low socioeconomic status were those below 25th percentile (or 10th percentile) of the SES distribution and high socioeconomic status are those above the 75th percentile (or 90th percentile) of the SES distribution. Details about the SES distribution and about the covariates are presented in the Methodology Section .

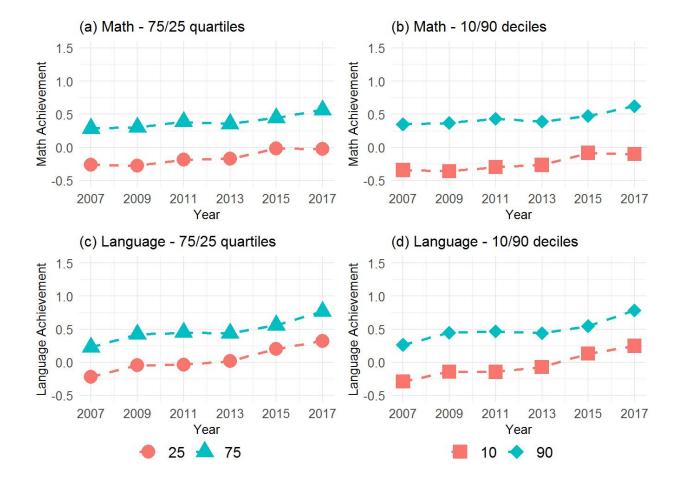


Figure 2. Brazil: 9th Grade Achievement Gap Trajectories, 2007-2017

Notes: This panel reports unconditional achievement gaps for language and math test scores between students in ninth grade considered high and low socioeconomic status (SES). Students considered low socioeconomic status were those below 25th percentile (or 10th percentile) of the SES distribution and high socioeconomic status are those above the 75th percentile (or 90th percentile) of the SES distribution. Details about the SES distribution and about the covariates are presented in the Methodology Section .

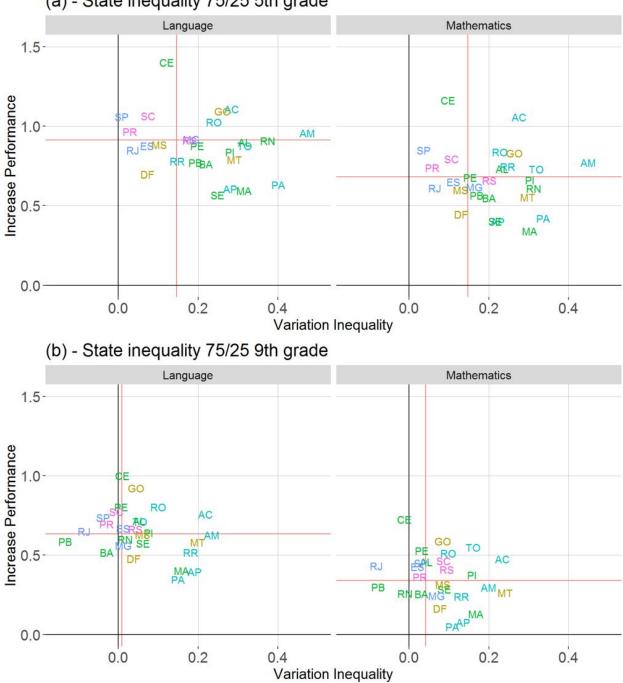


Figure 3. Brazil: Test score increases and SES achievement gap increases, by state (a) - State inequality 75/25 5th grade

Notes: This panel reports test score gains and achievement gaps for language and math test scores between students in fifth and ninth grades considered high and low socioeconomic status (SES). The gains for each state are represented by the states abbreviations while the red lines represent the national variations. Students considered low socioeconomic status were those below 25th percentile of the SES distribution and high socioeconomic status are those above the 75th percentile of the SES distribution. For this panel, we defined low and high SES within each state. Details about the SES distribution and about the covariates are presented in the Methodology Section.

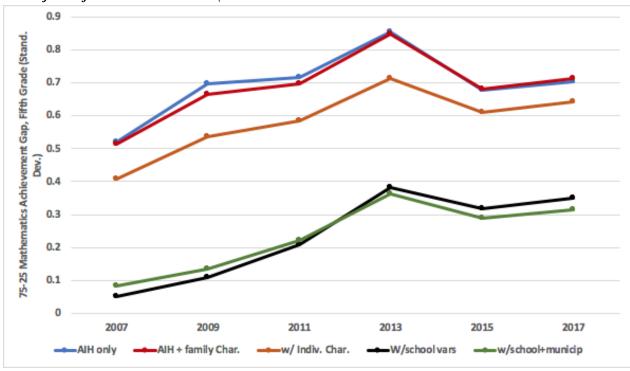
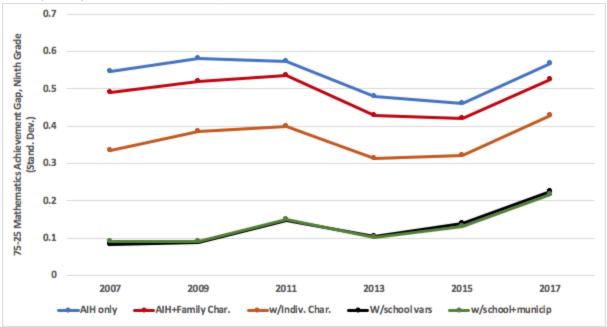


Figure 4. Brazil: Estimated 75-25 Mathematics Achievement Gaps 2007-2017, Fifth Grade, Unadjusted and Adjusted for Control Variables (standard deviations

Source: Table 2.

Figure 5. Brazil: Estimated 75-25 Mathematics Achievement Gaps 2007-2017, Ninth Grade, Unadjusted and Adjusted for Control Variables (standard deviations



Source: Table 4.

				GRA	DE 5				
			F est Scores ES	Gains Across Years SES					
Year	low	medium- low	medium- high	high	low	medium- low	medium- high	high	
2007	185.2	192.8	199.2	207.5	-	-	-	-	
2009	191.6	203.0	212.2	220.2	6.4	10.2	12.9	12.7	
2011	194.7	206.7	215.9	224.5	3.1	3.7	3.8	4.3	
2013	193.3	209.6	221.6	228.6	-1.5	2.8	5.7	4.1	
2015	202.9	216.2	225.8	231.9	9.6	6.6	4.2	3.3	
2017	206.5	220.4	230.1	236.9	3.7	4.2	4.3	5.1	
				GRA	DE 9				
			est Scores				ross Years		
			ES				ES		
		medium-	medium-			medium-	medium-		
Year	low	low	high	high	low	low	high	high	
2007	231.1	238.6	245.4	253.6	-	-	-	-	
2009	230.6	239.2	246.9	255.0	-0.5	0.6	1.5	1.4	
2011	233.9	243.4	250.8	257.8	3.3	4.2	3.9	2.9	
2013	233.7	243.9	251.0	256.6	-0.2	0.5	0.2	-1.2	
2015	240.4	248.9	255.3	260.7	6.8	5.0	4.3	4.1	
2017	240.4	250.6	259.0	266.1	0.0	1.7	3.7	5.5	

Table 1. Brazil: Mean test scores and gain across years, by grade and SES quartile, 2007-2017

Source: INEP, Prova Brasil.

Regression results: 5th grade										
				Dependen	t variable:					
				Langua	ge Score					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)		
Gap high and low SES (baseline)	0.483***	0.476***	0.389***	0.065***	0.048***	0.076***	0.083***	0.095***		
	(0.019)	(0.017)	(0.016)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)		
Additional gap - 2009	0.135***	0.107***	0.083***	0.010	0.012	0.011	0.010	0.008		
	(0.009)	(0.009)	(0.009)	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)		
Additional gap - 2011	0.137***	0.124***	0.114***	0.070***	0.092***	0.071***	0.070***	0.060***		
	(0.010)	(0.009)	(0.011)	(0.009)	(0.009)	(0.009)	(0.008)	(0.009)		
Additional gap - 2013	0.279***	0.280***	0.254***	0.242***	0.274***	0.227***	0.216***	0.198***		
	(0.016)	(0.015)	(0.015)	(0.012)	(0.011)	(0.014)	(0.011)	(0.010)		
Additional gap - 2015	0.179***	0.191***	0.221***	0.259***	0.281***	0.223***	0.209***	0.189***		
	(0.015)	(0.015)	(0.016)	(0.014)	(0.013)	(0.017)	(0.013)	(0.011)		
Additional gap - 2017	0.204***	0.220***	0.252***	0.291***	0.311***	0.246***	0.236***	0.214***		
	(0.015)	(0.015)	(0.016)	(0.014)	(0.013)	(0.016)	(0.014)	(0.011)		
Family features	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes		
Age, gender and race	No	No	Yes	Yes	Yes	Yes	Yes	Yes		
Average SES per school	No	No	No	Yes	Yes	Yes	Yes	Yes		
School and teacher features	No	No	No	No	Yes	Yes	Yes	Yes		
Municipality - economic features	No	No	No	No	No	Yes	Yes	Yes		
Municipality - policy features	No	No	No	No	No	No	Yes	Yes		
State fixed effect	No	No	No	No	No	No	No	Yes		
Observations	13,127,042	13,127,042	13,127,042	13,127,042	12,671,048	12,670,639	12,390,068	12,390,06		
Adjusted R ²	0.116	0.151	0.194	0.229	0.234	0.236	0.235	0.245		

Table 2. Brazil: Estimated 5th grade language scores with articles in the home measure of SES, 75/25 achievement gap

Notes: This table reports estimates of the achievement gaps for language test scores between students in fifth grade considered high and low socioeconomic status (SES). Students considered low socioeconomic status were those below 25th percentile of the SES distribution and high socioeconomic status are those above the 75th percentile of the SES distribution. Details about the SES distribution and about the covariates are presented in Section Methodology. Standard errors were clustered at the municipality level. *p<0.05; **p<0.01; **p<0.01

Regression results: 5th grade										
				Dependen	t variable:					
	-			Math	Score					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)		
Gap high and low SES (baseline)	0.521***	0.513***	0.407***	0.072***	0.052***	0.077***	0.083***	0.090***		
	(0.021)	(0.019)	(0.018)	(0.006)	(0.006)	(0.007)	(0.007)	(0.008)		
Additional gap - 2009	0.176***	0.152***	0.128***	0.052***	0.057***	0.055***	0.053***	0.050***		
	(0.010)	(0.010)	(0.009)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)		
Additional gap - 2011	0.196***	0.185***	0.177***	0.131***	0.158***	0.141***	0.138***	0.131***		
	(0.013)	(0.012)	(0.014)	(0.011)	(0.012)	(0.011)	(0.010)	(0.010)		
Additional gap - 2013	0.334***	0.336***	0.306***	0.293***	0.331***	0.290***	0.280***	0.267***		
	(0.018)	(0.017)	(0.017)	(0.014)	(0.013)	(0.015)	(0.012)	(0.011)		
Additional gap - 2015	0.156***	0.168 ^{***}	0.202***	0.241***	0.267***	0.218***	0.205***	0.189***		
	(0.016)	(0.016)	(0.018)	(0.016)	(0.015)	(0.018)	(0.015)	(0.013)		
Additional gap - 2017	0.184***	0.200***	0.237***	0.278***	0.297***	0.241***	0.232***	0.217***		
	(0.015)	(0.015)	(0.017)	(0.015)	(0.014)	(0.017)	(0.015)	(0.013)		
Family features	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes		
Age, gender and race	No	No	Yes	Yes	Yes	Yes	Yes	Yes		
Average SES per school	No	No	No	Yes	Yes	Yes	Yes	Yes		
School and teacher features	No	No	No	No	Yes	Yes	Yes	Yes		
Municipality - economic features	No	No	No	No	No	Yes	Yes	Yes		
Municipality - policy features	No	No	No	No	No	No	Yes	Yes		
State fixed effect	No	No	No	No	No	No	No	Yes		
Observations	13,127,042	13,127,042	13,127,042	13,127,042	12,671,048	12,670,639	12,390,068	12,390,06		
Adjusted R ²	0.098	0.129	0.166	0.207	0.216	0.217	0.216	0.227		

Table 3. Brazil: Estimated 5th grade mathematics scores with articles in the home measure of SES, 75/25 achievement gap

Note: This table reports estimates of the achievement gaps for math test scores between students in fifth grade considered high and low socioeconomic status (SES). Students considered low socioeconomic status were those below 25th percentile of the SES distribution and high socioeconomic status are those above the 75th percentile of the SES distribution. Details about the SES distribution and about the covariates are presented in Section Methodology. Standard errors were clustered at the municipality level. p<0.05; p<0.01; p<0.01

Table 4. Brazil: Estimated 9th grade language scores with articles in the home measure of SES, 75/25 achievement gap

Regression results: 9th grade										
	6			Dependen	t variable:					
				Langua	ge Score					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)		
Gap high and low SES (baseline)	0.444***	0.396***	0.314***	0.102***	0.084***	0.093***	0.097***	0.113***		
	(0.015)	(0.013)	(0.011)	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)		
Additional gap - 2009	0.025**	0.020*	0.040***	-0.010	-0.003	-0.004	-0.006	-0.015*		
	(0.009)	(0.009)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)		
Additional gap - 2011	0.046***	0.065***	0.084***	0.053***	0.078***	0.073***	0.070***	0.053***		
	(0.008)	(0.008)	(0.008)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)		
Additional gap - 2013	-0.067***	-0.060***	-0.022*	-0.018*	0.014	-0.001	-0.006	-0.027***		
	(0.009)	(0.008)	(0.009)	(0.008)	(0.008)	(0.009)	(0.009)	(0.008)		
Additional gap - 2015	-0.085***	-0.071***	-0.018	0.016	0.043***	0.025*	0.014	-0.009		
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.012)	(0.011)	(0.009)		
Additional gap - 2017	-0.010	0.002	0.052***	0.071***	0.095***	0.075***	0.067***	0.040***		
	(0.016)	(0.015)	(0.014)	(0.012)	(0.012)	(0.016)	(0.014)	(0.011)		
Family features	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes		
Age, gender and race	No	No	Yes	Yes	Yes	Yes	Yes	Yes		
Average SES per school	No	No	No	Yes	Yes	Yes	Yes	Yes		
School and teacher features	No	No	No	No	Yes	Yes	Yes	Yes		
Municipality - economic features	No	No	No	No	No	Yes	Yes	Yes		
Municipality - policy features	No	No	No	No	No	No	Yes	Yes		
State fixed effect	No	No	No	No	No	No	No	Yes		
Observations	10,855,470	10,855,470	10,855,470	10,855,470	10,512,931	10,512,273	10,310,264	10,310,264		
Adjusted R ²	0.049	0.076	0.129	0.146	0.152	0.152	0.152	0.162		

Note: This table reports estimates of the achievement gaps for language test scores between students in ninth grade considered high and low socioeconomic status (SES). Students considered low socioeconomic status were those below 25th percentile of the SES distribution and high socioeconomic status are those above the 75th percentile of the SES distribution. Details about the SES distribution and about the covariates are presented in Section Methodology. Standard errors were clustered at the municipality level. p<0.05; p<0.01; p<0.01

				Dependen	t variable:					
	Math Score									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)		
Gap high and low SES (baseline)	0.548***	0.490***	0.336***	0.107***	0.084***	0.085***	0.090***	0.111***		
	(0.021)	(0.019)	(0.016)	(0.006)	(0.006)	(0.009)	(0.008)	(0.007)		
Additional gap - 2009	0.034***	0.031***	0.050***	-0.005	0.004	0.003	0.002	-0.013		
	(0.009)	(0.009)	(0.010)	(0.008)	(0.008)	(0.008)	(0.007)	(0.008)		
Additional gap - 2011	0.026**	0.045***	0.062***	0.028**	0.062***	0.062***	0.060***	0.037***		
	(0.010)	(0.009)	(0.010)	(0.009)	(0.009)	(0.007)	(0.008)	(0.009)		
Additional gap - 2013	-0.069***	-0.060***	-0.023**	-0.020*	0.021**	0.019	0.012	-0.017		
	(0.009)	(0.009)	(0.009)	(0.008)	(0.008)	(0.011)	(0.010)	(0.009)		
Additional gap - 2015	-0.086***	-0.070***	-0.015	0.021*	0.056***	0.054***	0.041***	0.011		
	(0.010)	(0.010)	(0.009)	(0.009)	(0.009)	(0.014)	(0.012)	(0.010)		
Additional gap - 2017	0.021	0.035	0.094***	0.114***	0.142***	0.139***	0.127***	0.089***		
	(0.019)	(0.018)	(0.016)	(0.013)	(0.014)	(0.021)	(0.017)	(0.013)		
Family features	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes		
Age, gender and race	No	No	Yes	Yes	Yes	Yes	Yes	Yes		
Average SES per school	No	No	No	Yes	Yes	Yes	Yes	Yes		
School and teacher features	No	No	No	No	Yes	Yes	Yes	Yes		
Municipality - economic features	No	No	No	No	No	Yes	Yes	Yes		
Municipality - policy features	No	No	No	No	No	No	Yes	Yes		
State fixed effect	No	No	No	No	No	No	No	Yes		
Observations	10,855,470	10,855,470	10,855,470	10,855,470	10,512,931	10,512,273	10,310,264	10,310,26		
Adjusted R ²	0.046	0.069	0.111	0.131	0.140	0.141	0.141	0.154		

Table 5. Brazil: Estimated 9th grade mathematics scores with articles in the home measure of SES, 75/25 achievement gap

Note: This table reports estimates of the achievement gaps for math test scores between students in ninth grade considered high and low socioeconomic status (SES). Students considered low socioeconomic status were those below 25th percentile of the SES distribution and high socioeconomic status are those above the 75th percentile of the SES distribution. Details about the SES distribution and about the covariates are presented in Section Methodology. Standard errors were clustered at the municipality level. *p<0.05; **p<0.01; ***p<.001

Regression results: 5th grade												
	22	Dependent variable:										
	Language Score				Math Score							
	SES: AIH	SES: FULL	ES: FULL SES: AIH	SES: FULL	SES: AIH	SES: FULL	SES: AIH	SES: FULL				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)				
Gap high and low SES (baseline)	0.483***	0.501***	0.095***	0.051***	0.521***	0.536***	0.090***	0.056***				
	(0.019)	(0.019)	(0.007)	(0.008)	(0.021)	(0.020)	(0.008)	(0.008)				
Additional gap - 2009	0.135***	0.125***	0.008	0.011	0.176***	0.166***	0.050***	0.052***				
	(0.009)	(0.010)	(0.007)	(0.007)	(0.010)	(0.010)	(0.007)	(0.007)				
Additional gap - 2011	0.137***	0.126***	0.060***	0.057***	0.196***	0.189***	0.131***	0.130****				
	(0.010)	(0.010)	(0.009)	(0.009)	(0.013)	(0.013)	(0.010)	(0.011)				
Additional gap - 2013	0.279***	0.279***	0.198***	0.189***	0.334***	0.332***	0.267***	0.257***				
	(0.016)	(0.015)	(0.010)	(0.010)	(0.018)	(0.017)	(0.011)	(0.011)				
Additional gap - 2015	0.179***	0.176***	0.189***	0.183***	0.156***	0.154***	0.189***	0.183***				
	(0.015)	(0.014)	(0.011)	(0.010)	(0.016)	(0.015)	(0.013)	(0.012)				
Additional gap - 2017	0.204***	0.180***	0.214***	0.197***	0.184***	0.164***	0.217***	0.204***				
	(0.015)	(0.015)	(0.011)	(0.011)	(0.015)	(0.015)	(0.013)	(0.012)				
Family features	No	No	Yes	-	No	No	Yes	-				
Age, gender and race	No	No	Yes	Yes	No	No	Yes	Yes				
Average SES per school	No	No	Yes	Yes	No	No	Yes	Yes				
School and teacher features	No	No	Yes	Yes	No	No	Yes	Yes				
Municipality - economic features	No	No	Yes	Yes	No	No	Yes	Yes				
Municipality - policy features	No	No	Yes	Yes	No	No	Yes	Yes				
State fixed effect	No	No	Yes	Yes	No	No	Yes	Yes				
Observations	13,127,042	13,088,535	12,390,068	12,353,309	13,127,042	13,088,535	12,390,068	12,353,309				
Adjusted R ²	0.116	0.117	0.245	0.231	0.098	0.099	0.227	0.216				

Table 6. Brazil: Estimated 5th grade scores with both versions of the PCA measure of SES (articles in the home versus full version), 75/25 achievement gap

Note: This table reports estimates of the achievement gaps for between fifth grade students in high and low socioeconomic status (SES). We compare two different SES definitions. The AIH is the same SES distribution than used in Tables 1-4. The FULL definition includes family characteristics in the PCA. More details are presented in Section Methodology. Standard errors were clustered at the municipality level. *p<0.05; **p<0.01; **p<0.01

	Dependent variable:									
	6	Langua	ge Score		Math Score					
	SES: AIH	SES: FULL	SES: AIH	SES: FULL	SES: AIH	SES: FULL	SES: AIH	SES: FULL		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)		
Gap high and low SES (baseline)	0.444***	0.458***	0.113***	0.098***	0.548***	0.567***	0.111***	0.110***		
	(0.015)	(0.015)	(0.006)	(0.006)	(0.021)	(0.021)	(0.007)	(0.007)		
Additional gap - 2009	0.025**	0.030***	-0.015*	-0.010	0.034***	0.037***	-0.013	-0.011		
	(0.009)	(0.008)	(0.007)	(0.007)	(0.009)	(0.009)	(0.008)	(0.008)		
Additional gap - 2011	0.046***	0.056***	0.053***	0.047***	0.026**	0.034**	0.037***	0.032***		
	(0.008)	(0.009)	(0.007)	(0.008)	(0.010)	(0.010)	(0.009)	(0.009)		
Additional gap - 2013	-0.067***	-0.060***	-0.027***	-0.029***	-0.069***	-0.064***	-0.017	-0.022*		
	(0.009)	(0.009)	(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)		
Additional gap - 2015	-0.085***	-0.084***	-0.009	-0.012	-0.086***	-0.089***	0.011	0.004		
51	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)		
Additional gap - 2017	-0.010	-0.003	0.040***	0.036**	0.021	0.025	0.089***	0.085***		
	(0.016)	(0.016)	(0.011)	(0.011)	(0.019)	(0.019)	(0.013)	(0.013)		
Family features	No	No	Yes	-	No	No	Yes	-		
Age, gender and race	No	No	Yes	Yes	No	No	Yes	Yes		
Average SES per school	No	No	Yes	Yes	No	No	Yes	Yes		
School and teacher features	No	No	Yes	Yes	No	No	Yes	Yes		
Municipality - economic features	No	No	Yes	Yes	No	No	Yes	Yes		
Municipality - policy features	No	No	Yes	Yes	No	No	Yes	Yes		
State fixed effect	No	No	Yes	Yes	No	No	Yes	Yes		
Observations	10,855,470	10,855,065	10,310,264	10,309,891	10,855,470	10,855,065	10,310,264	10,309,891		
Adjusted R ²	0.049	0.052	0.162	0.151	0.046	0.049	0.154	0.147		

Table 7. Brazil:Estimated 9th grade scores with both versions of the PCA measure of SES (articles in the home versus full version), 75/25 achievement gap

Note: This table reports estimates of the achievement gaps for between ninth grade students in high and low socioeconomic status (SES). We compare two different SES definitions. The AIH is the same SES distribution than used in Tables 1-4. The FULL definition includes family characteristics in the PCA. More details are presented in Section Methodology. Standard errors were clustered at the municipality level. *p<0.05; **p<0.01; **p<0.01

Appendix A – Missing Values and Imputation Analysis

In order to apply PCA on the derivation of the SES indicator, it is necessary to impute missing values or remove the observations that have missing values from the original dataset. Before considering the impact of different imputation methods, we performed an initial analysis of the problem by characterizing the frequency of missing values for each variable and where they are concentrated. The variables included in the construction of the SES indicator were extracted from the SAEB student survey and can be broadly classified as articles in the home, parents' education and family structure variables. For the chosen SES, we include only variables related to articles in the home for the PCA. For the full version of the PCA (also tested), the complete list of variables is presented in Table A-1.

Variable	Name	Туре	Description
Bedrooms	student_bedrooms	Categorical	Number of bedrooms in the house
Bathrooms	student_bathroom	Categorical	Number of bathrooms in the house
Cars	student_car	Categorical	Number of cars in house

Table A-1. Brazil: PCA variables

Computers	student_computer	Boolean	Presence of computers in the house
Fridge	student_fridge	Categorical	Number of fridges in the house
Maid	student_maid	Categorical	Number/frequency of maids in the house
Mother Presence	student_living_mother	Categorical	Student lives with the mother
Father Presence	student_living_father	Categorical	Student lives with the father
Radios	student_radio	Categorical	Number of radios in the house
Televisions	student_tv	Categorical	Number of televisions in the home
DVD Player	student_video_dvd	Boolean	Presence of DVD players in the home
Washing Machine	student_washing_mach	Boolean	Presence washing machines in the home

Housemates	student_housemates	Categorical	Number of people living in the home
Literate Mother	student_mother_literate	Boolean	Mother is literate
Literate Father	student_father_literate	Boolean	Father is literate

In the analysis that follows, we first separated our dataset into two groups, that we call Core and Removed. The first corresponds to the students that have no missing values for any of the PCA variables and the mathematics and language scores. The second group is that of the students initially removed from the analysis, either for presenting missing values in the PCA variables or in at least one of the test scores. To help us understand the dimension of the problem of missing values, Table A-2 shows the distribution of the data between the two groups described previously, for each year.

Table A-2. Brazil: Percentage of missing values

Year	Students with Math Score [%]	Students with Language Score [%]	Students with All SES Variables [%]	Students with All SES Variables and Both Scores [%]	Number of Observations
2007	99.86	99.86	66.04	66	4,115,190
2009	76.37	76.39	48.59	48.54	5,931,024
2011	82.04	82.04	56.92	56.92	5,201,730
2013	78.45	78.45	59.91	59.66	5,244,713
2015	79.8	79.8	65.45	65.42	4,916,807
2017	80.55	80.55	62.66	62.64	4,965,478

In the table, we see that, for most years, the Core sample represents around 60% of the original dataset, with this value going as low as 48.54% for 2009. By using imputation methods (explored next), we are able to recover a portion of the data that is given by the difference between the minimum value in the second and third columns (which are usually equal) and the fourth column. This sample represents the students that have scores reported but show missing values in some of the variables used in the SES indicator. Roughly, we see that this value varies from 15% to 30%, approximately. At the same time that these large figures make the problem more relevant, they also make it harder to solve since it affects a significant portion of students that are not represented in the Core sample. For this reason, any conclusions derived from a model that excludes observations with missing values cannot be easily extrapolated to the entire dataset.

In addition to the dimension of the data loss due to missing values, it is important to evaluate the possibility of bias being introduced to the dataset when we exclude students. We compared basic statistics related to each variable in the Core and the Removed groups. The results are presented in Table A-3, which also includes the p-value for the two-sided t-test of the difference of the statistics in the two samples.

Variable	Core Group	Students Removed from Analysis	P-Value
2 or more bedrooms [%]	88.89	84.51	0
2 or more bathrooms [%]	29.1	25.38	0
At least 1 car [%]	48.73	40.8	0
At least 1 computer [%]	53.62	43.91	0
At least 1 fridge [%]	97.51	94.27	0

 Table A-3. Brazil: Descriptive Analysis of the proportion of students with the indicated value of the article in the Core versus Removed Groups

At least 1 maid [%]	10.54	12.41	0
Living with the mother [%]	90.15	87.05	0
Living with the father [%]	66.2	59.38	0
At least 1 radio [%]	80.9	81.38	0
2 or more televisions [%]	55.69	50.73	0
At least 1 DVD player [%]	83.64	79.07	0
At least 1 washing machine [%]	72.94	66.81	0
5 or more people in the house [%]	50.04	56.82	0
Mother is literate [%]	94.67	92.56	0
Father is literate [%]	90.85	88.15	0
Mean math score - 5th grade	213.57	193.73	0
Mean math score - 9th grade	247.68	237.99	0
Mean language score - 5th grade	198.24	177.02	0
Mean language score - 9th grade	243.34	232.31	0
Number of observations	18,011,577	12,363,365	

The p-values observed on the last column are all very small due to the large size of the datasets. Despite this, some of the differences observed in the table, although significant, are only marginal, as is the case of the percentage of students with at least one radio at home. Generally, it seems that the Core sample consists of more privileged students in comparison to the ones

removed from analysis. Some of the most important statistics such as the percentage of students living with their fathers, the number of people in the house, the presence of a computer, and even the test scores, indicate the same conclusion. We also observe that the students with missing values perform worse than the Core group. These observations combined suggest that we might be underestimating the achievement gap. If we could include all students in the analysis, it is likely that we would be adding students in the lowest quartiles/deciles of the SES measure that would lead to a lower academic achievement at the bottom of the distribution.

Given the conclusions exposed previously, we explored different imputation strategies to solve the problem of missing values. We tested replacing these values with the mode, maximum, minimum and a random selection for each variable. In all these strategies the replacement was made at the school level. The last three methods (maximum, minimum and random) are not expected to be reliable solutions. Instead, we use these extreme methods of imputation to assess the impact of different strategies on the results. Figures A-1 and A-2 show the estimated gap for each year using the different imputation strategies.

Figure A-1: Brazil: Achievement Gaps (25/75)--5th Grade Math, Imputation Comparison

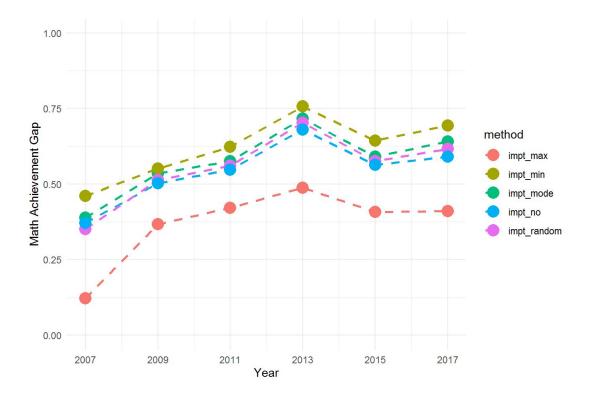
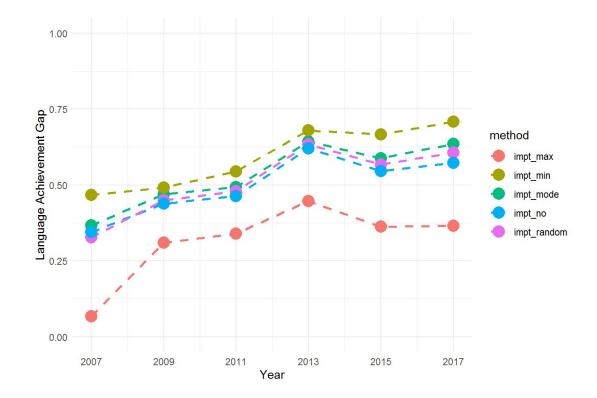


Figure A-2: Brazil: Achievement Gaps (25/75)--5th Grade Language, Imputation Comparison



We can draw interesting conclusions from these results. First, our initial hypothesis of an underestimation of the gap caused by the missing values seems to be correct. We observe that the curve representing no imputation (impt_no) generally has lower values than the curves representing the minimum (impt_min), mode (impt_mode), and random (impt_random) imputation. Only the maximum imputation generates lower gaps than the dataset with missing values. This is because, as our previous analysis showed, the students with missing values tend to be at the bottom of the SES distribution. Therefore, by using the maximum (wealthiest) values within school, we artificially move up in the SES distribution students that were originally at the bottom of both the SES and scores distributions. On the other hand, the minimum imputation does not seem to have as big of an impact as the maximum since the dislocation of students within the distributions is much smaller (they were already close to the bottom).

The previous observations suggest the adoption of an imputation strategy in order to avoid the distortions introduced by the removal of missing values. A common strategy is the multivariate imputation by chained equations (MICE). The idea is to use all other variables in the dataset to predict the value of those that are missing. One model is created for each variable and the process is repeated for all variables until convergence or a predefined stopping criterion is met. This method can handle statistical uncertainty better than simple strategies such as mean or mode imputation. However, it would demand intense computational effort for our very large dataset. Additionally, as we can observe in Figures A-1 and A-2, there is only a marginal variation on the results for the curves representing random, mode, and no imputation. In other words, the computational effort of implementing a more sophisticated method would most likely not be compensated by an improvement on results. For this reason, we decided to use the within-school mode imputation mentioned previously.

Appendix B – PCA Computation and Evaluation

In this section we evaluate the SES index generated through PCA. The main challenge is to choose a group of variables that are strongly correlated with the socioeconomic level of students to include in the construction of the index. This group has to be representative enough to capture the influence of SES and educational achievements but it has to be small enough to allow an effective dimensionality reduction while preserving information. We tested two different groups of variables for the PCA, one that includes only variables related to articles in the home and another that also includes parent's education and family structure. These include student_maid, student_living_mother, student_living_father, student_housemates, student_mother_literate, and student_father_literate. All other variables shown in Table A-1 were included in both the full and articles in the home (AIH) versions of PCA.

In order to effectively preserve the information contained in multiple dimensions after the dimensionality reduction with PCA, the variables need to have a strong enough correlation. We can initially assess this aspect through a correlogram of the variables, shown in Figures B-1 and B-2.

As we can observe in the correlograms, the most strongly correlated variables seem to be the ones that represent articles in the home. For example, student_bathroom is strongly correlated with student_bedrooms, student_car, student_fridge, and student_tv. This observation was the main motivator to test the PCA with only articles in the home. Figures B-3 and B-4 show the weights of each variable in both versions of the PCA (the principal components are weighted linear combinations of the variables).

Since our SES index consists of the first principal component (PC1), we would want all variables to receive a significant weight for this component. For the AIH-PCA, we see that except for radio and video_dvd, all variables receive similarly high weights (we are concerned with the absolute values). However, for the FULL-PCA most variables related to parent's education (mother_literate, father_literate) and household structure (living_mother, living_father, housemates, maid) receive a low weight in the first component. These variables receive a higher weight for the second component, meaning that the variance present in the original dataset

associated with these variables is captured by the second component, not the first. For this reason, in the FULL-PCA version of the SES measure, the variables not related to articles in the home are effectively being excluded from the index. This is a strong reason to choose the AIH-PCA over FULL-PCA. Another reason is summarized in Figures B-5 and B-6, that show the portion of the original variance explained by each principal component.

The variance explained by the first component in the FULL-PCA is around 0.2 while in the AIH-PCA it is around 0.3. This means that the SES index that uses only articles in the home preserves better the information contained in the original dataset after the dimensionality reduction. This increases our confidence in the index and its capacity of representing the socioeconomic level of students. In table B-1, we can further explore the correlation between our SES index and the real socioeconomic level of students, for the case of the AIH-PCA.

Table B1 shows the same statistics presented in Table A-3 but separating students according to their position in the SES index distribution. It is interesting to note that our measure is indeed capable of capturing relevant differences between the groups of students. For example, if we compare the first and last columns, relevant statistics such as the ones associated with the number of bedrooms and bathrooms, the presence of computers, the number of televisions and the literacy of the parents are remarkably different. The scores are also increasing monotonically as we move towards the upper groups of the SES distribution, demonstrating once again the correlation between SES and educational performance. We also observe that throughout the entire range of the SES measure, there are consistent trends for the statistics. This suggests that the correlation (or at least its sign) between variables is also consistent throughout the same range. If this were not the case, it would be hard for the PCA to effectively reduce the dimensionality of the data while preserving information.

As a last evaluation of the quality of our SES index, we compare it with the Nivel Socioeconomico (NSE) estimated by the Insituto Nacional de Estudos e Pesquisas Educacionais Anisio Teixeira (INEP). This is an official SES measure constructed on the school-level. To compare both measures, we aggregate our index capturing the mean per school. Figures B-7 and B-8 show the data per school with a regression line (blue line) and the local mean (yellow points).

It is clear that both measures are strongly correlated. There is a clear linear relationship and the line intercepts the origin. The slope is close to 1, suggesting that both measures capture very similar patterns. This result once again increases our confidence on the constructed SES index. This analysis shows that it is capable of preserving the information contained in the original multidimensional dataset in a way that is consistent with externally observed socioeconomic trends.

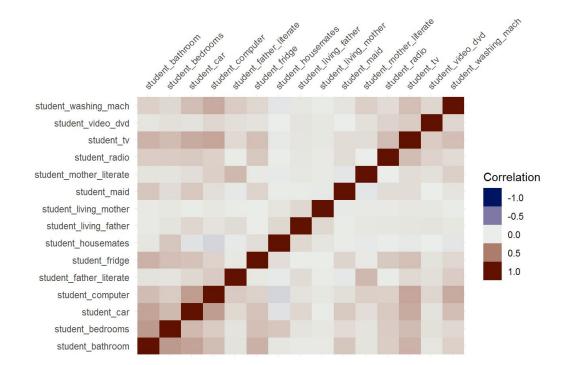


Figure B-1: Correlogram PCA Variables - 2015 (5th grade)

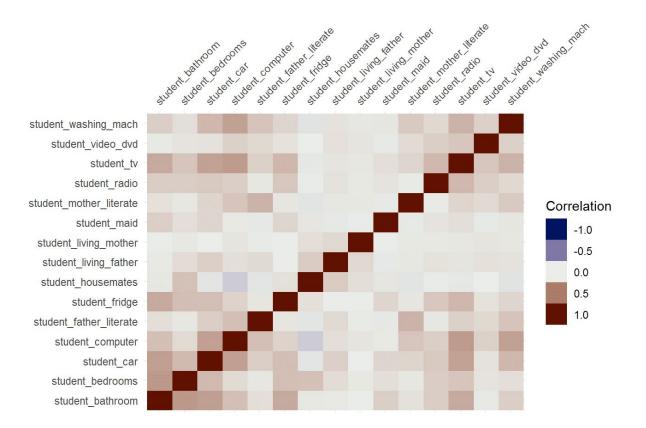


Figure B-2: Correlogram PCA Variables - 2015 (9th grade)

Figure B-3: Factor loading of AIH-PCA variables for the first and second components - 2015 (5th grade)

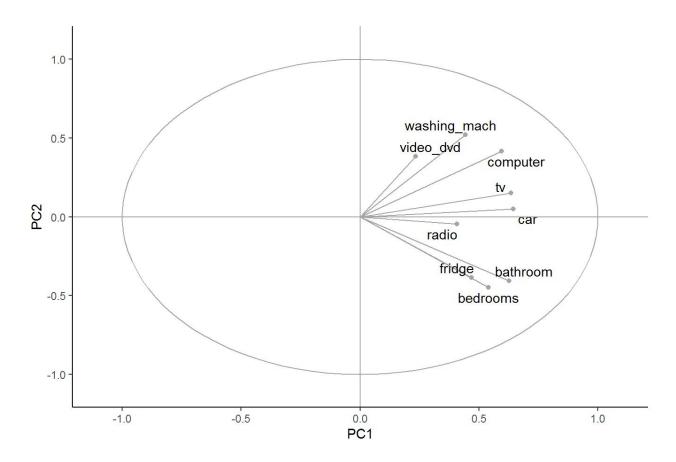


Figure B-4: Factor loading of FULL-PCA variables for the first and second components - 2015 (5th grade)

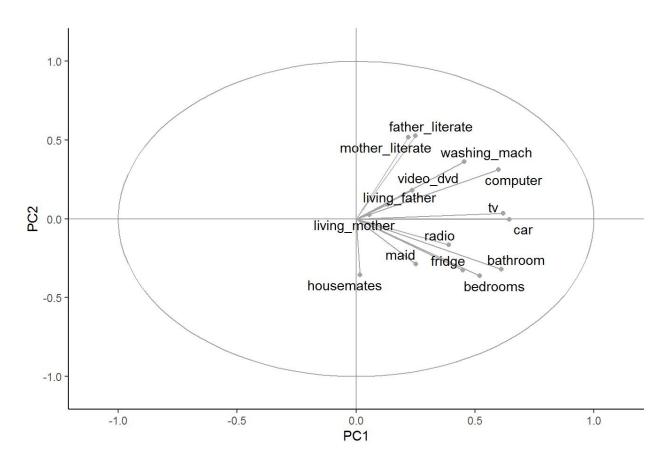


Figure B-5: Variance explained by each principal component of the AIH-PCA - 2015 (5th grade)

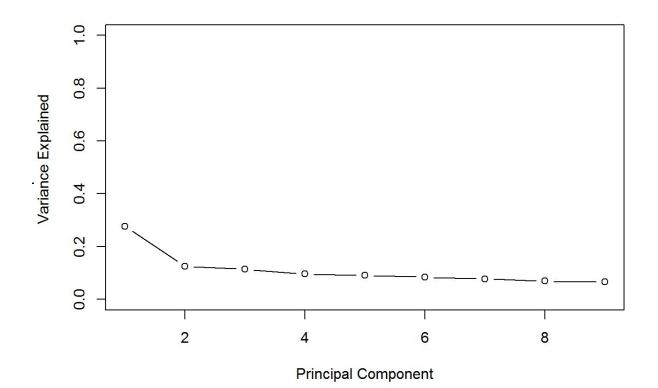
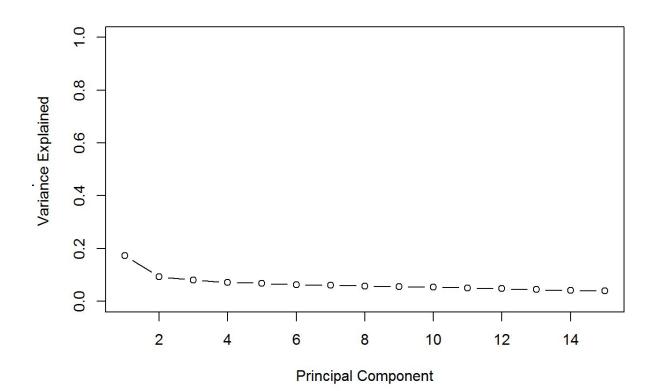


Figure B-6: Variance explained by each principal component of the FULL-PCA - 2015 (5th grade)



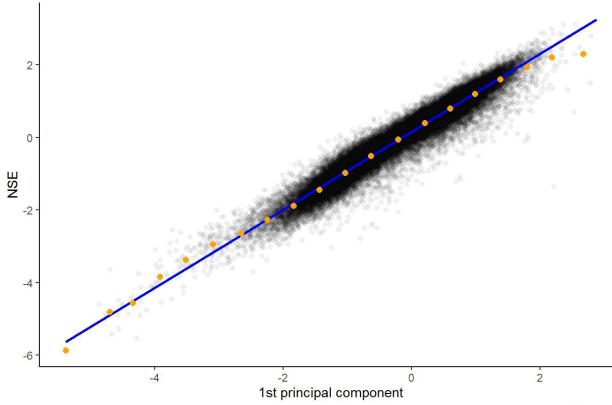


Figure B-7: Correlation between NSE and the first principal component - 2015 (5th grade)

slope = 1.1

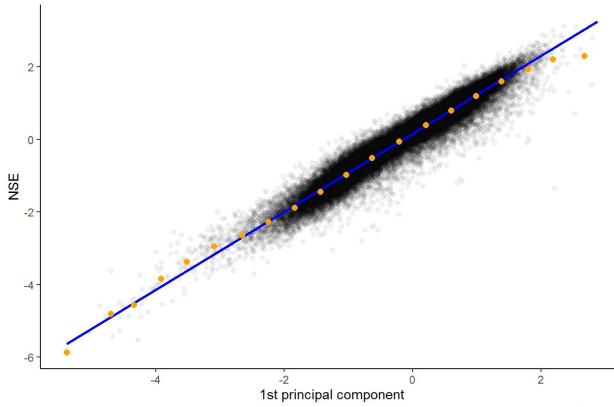


Figure B-8: Correlation between NSE and the first principal component - 2015 (9th grade)

slope = 1.1

Variable	Percentile 10	Percentile 10-25	Percentile 25-50	Percentile 50-75	Percentile 75-90	Percentile 90-100
2 or more bedrooms [%]	64.27	79.19	88.16	95.23	98.54	99.52
2 or more bathrooms [%]	1.32	4.62	12.69	29.66	59.07	88.29
At least 1 car [%]	4.3	14.33	34.95	63.39	81.67	93.2
At least 1 computer [%]	3.75	14.49	42.13	71.85	87.22	94.92
At least 1 fridge [%]	82.32	97.66	99.13	99.58	99.75	99.88
At least 1 maid [%]	5.92	6.35	7.45	9.42	14.02	26.78
Living with the mother [%]	88	89.71	90.55	91.12	90.66	88.78
Living with the father [%]	57.09	59.89	63.73	69.39	72.79	73.1
At least 1 radio [%]	55.59	71.84	79.52	86.01	91.46	94.63
2 or more televisions [%]	5.26	18.69	44.12	72.94	89.39	96.88
At least 1 DVD player [%]	48.1	74.98	84.61	90.82	94.11	96.08
At least 1 washing machine [%]	16.77	45.56	73.27	88.48	94.76	97.72
5 or more people in the house [%]	53.89	52.1	49.04	46.77	48.67	55.84
Mother is literate [%]	84.25	90.84	94.88	97.25	98.18	98.59

Table B-1: Socioeconomic statistics for different groups of students along the SES indexdistribution - 2015

Father is literate [%]	75.96	84.54	90.82	94.76	96.43	97.15
Mean math score - 5th grade	191.67	201.06	210.13	219.84	225.86	228.72
Mean language score - 5th grade	178.37	187.48	195.7	204.27	208.75	209.81
Mean language score - 9th grade	229.22	235.42	241.36	247.6	251.13	251.95
Mean math score - 9th grade	231.42	237.73	244.56	252.08	257.51	260.87
Number of observations	1,801,158	2,701,737	4,502,894	4,502,894	2,701,737	1,801,157

Appendix C - List of auxiliary variables

In this section we present the municipality and school-level control variables used in the main regression analysis. Table C-1 shows the municipality-level features.

Variable	Name	Description
GDP per capita	mun_gdp	Gross domestic product per capita in the municipality
Education spending	mun_exp_edu	Municipality annual expenditure on education and culture
Health spending	mun_exp_health	Municipality annual expenditure on health and sanitation
Transportation spending	mun_exp_trans	Municipality annual expenditure on public transportation
Public housing spending	mun_exp_house	Municipality annual expenditure on public housing
Welfare spending	mun_exp_welf	Municipality annual expenditure on assistance and welfare
Bolsa Familia spending	mun_bolsa_value	Annual Bolsa Familia (a federal government cash transfer program) spending

 Table C-1 . Municipality-level variables.

In addition to the municipality features, we also included school-level variables in our analysis. These represent teachers' behaviors, education, and the average SES measure for the school. The original dataset included many features (76) related to teachers. In order to select only those that are correlated with the educational achievement of students, we first use a LASSO model for feature selection. In this analysis, we use the mean score per school as the dependent variable and use all school-level features as covariates. The optimal lambda (regularization parameter) was chosen using a standard cross-validation procedure. The selected features are listed in Table C2.

Variable	Name	Description
Average SES	sch_ses	Average SES measure of students in the school
Average coverage of curriculum	teacher_content_achievement	Average percentage of curriculum content covered by teachers in the school
Average teachers' age	teacher_age	Average age of teachers in the school
Teachers constrained by students behaviour	teacher_constrain_behaviour	Percentage of teachers that have experienced disruptive students' behaviour
Teachers report drug use among students	teacher_drugs	Percentage of teachers that have observed students going to school under the effect of drugs
Education level of the teachers	teacher_education	Percentage of teachers that have completed a higher education degree
Most common race among teachers	teacher race	Mode of teacher's race
Access to necessary tools	teacher_tools_copier	Percentage of teachers that have used a photocopier in the school