NBER WORKING PAPER SERIES

SCHOOL BUS EMISSIONS, STUDENT HEALTH, AND ACADEMIC PERFORMANCE

Wes Austin
Garth Heutel
Daniel Kreisman

Working Paper 25641
http://www.nber.org/papers/w25641

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2019

We thank the GaDER program for assistance in identifying retrofits. We thank Jonathan Smith, Ariell Zimran, and seminar participants at TEAM-Fest, the Southern Economics Association annual meeting, and the University of South Carolina for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

School Bus Emissions, Student Health, and Academic Performance
Wes Austin, Garth Heutel, and Daniel Kreisman
NBER Working Paper No. 25641
March 2019
JEL No. I18,I20,Q53

ABSTRACT

Diesel emissions from school buses expose children to high levels of air pollution; retrofitting bus engines can substantially reduce this exposure. Using variation from 2,656 retrofits across Georgia, we estimate effects of emissions reductions on district-level health and academic achievement. We demonstrate positive effects on respiratory health, measured by a statewide test of aerobic capacity. Placebo tests on body mass index show no impact. We also find that retrofitting districts see significant test score gains in English and smaller gains in math. Results suggest that engine retrofits can have meaningful and cost-effective impacts on health and cognitive functioning.

Wes Austin
Georgia State University
gaustin4@gsu.edu

Garth Heutel
436 Andrew Young School
Department of Economics
Georgia State University
PO Box 3992
Atlanta, GA 30302-3992
and NBER
gheutel@gsu.edu

Daniel Kreisman
Department of Economics
Andrew Young School of Policy Studies
P.O. Box 3992
Atlanta, GA 30302-3992
Georgia
dkreisman@gsu.edu

1 Introduction

Nearly 25 million children ride over 500,000 buses to school in the United States each day. The predominantly diesel bus fleet contributes to air pollution exposure that may adversely affect children's health and academic performance. Because of this, school bus retrofit programs have been enacted across the country, making use of up to $200 million in federal grants per year to local districts to replace or retrofit engines. We use information on 2,656 of these school bus retrofits in Georgia, affecting approximately 150,000 students, to estimate effects on student health and academic achievement.

Diesel retrofits are an immediate and relatively inexpensive way to dramatically reduce diesel emissions.[1] A large literature has estimated the effect of diesel engine emissions on ambient air quality, in particular on nitrogen oxide and particulate matter.[2] A separate literature examines the effect of exposure to air pollution on children's academic achievement and health.[3] Yet, little is known about the direct effect of diesel emission reductions on children's academic achievement or health. The only studies to investigate school bus retrofits on health outcomes are Beatty and Shimshack (2011), which finds that bus retrofits in Washington state lead to significant reductions in asthma and pneumonia doctor visits, and Adar et al. (2015), which finds that retrofits in Washington state reduce pollution and pulmonary inflammation and increased lung growth. No study we know of examines the effect of reduced exposure to school bus emissions on academic performance.

To address the causal link between diesel retrofits, student health and academic achievement, we exploit variation in the timing and location of over 2,600 school bus retrofits across Georgia between 2007 and 2015. During our sample period, 15 percent of Georgia's 180 school districts retrofitted a share of their fleet. Our measure of exposure at the district level is based on the proportion of the bus fleet retrofitted in a given district. We further refine this with the proportion of students who are bus riders and the average amount of time students spend on the bus. We match retrofitting data to two types of district-level outcome measures: student health and scholastic outcomes. For the former, we observe a state-mandated fitness evaluation known as FitnessGram.[4] These health

[1] Barone et al. (2010); Tate et al. (2017).
[2] EPA (2003).
[3] Currie and Neidell (2005); Lavy et al. (2014).
[4] The FitnessGram© tests have been used for decades to assess student health, and a large literature demonstrates the scientific validity of the tests employed. The FitnessGram manual (https://www.cooperinstitute.org/vault/2440/web/files/662.pdf) provides details.

data include an established measure of cardiovascular health (aerobic capacity), which allows us to estimate effects on respiratory health, and BMI, which we take as a potential placebo against general health trends, though we discuss why BMI might also be affected by improved respiratory health. For scholastic outcomes we observe English and math end-of-grade test scores, in addition to attendance.

We find positive and non-trivial effects of bus retrofits on student health. Retrofitting an entire fleet leads to a 4 percent increase in that district's average aerobic capacity, or roughly 1.8 units of VO_2 max, in our most conservative estimate. This effect is slightly larger when we weight treatment by the share of students in a district who ride the bus. In this case, retrofitting 100 percent of buses in a district where everyone rides the bus would yield a 5 percent improvement in aerobic capacity. We find no relationship between retrofits and our placebo, BMI. We show that effects on aerobic capacity are strongest for elementary school students.

We also find evidence that these retrofits affected student achievement. A retrofit of 10 percent of a district's fleet increases English test scores by 0.009 standard deviations, so retrofitting an entire district's fleet would increase test scores by nearly one-tenth of a standard deviation. Weighting by the share of students who ride the bus, we estimate that districts would see a 0.14 standard deviation increase from retrofitting an entire fleet when all students ride the bus. Estimated effects on math scores are also positive, but are smaller and noisier than those for English and often cannot be distinguished from zero. We find little evidence that attendance was significantly affected, though initial attendance rates were very high.

Our results suggest that retrofits are a cost-effective lever to improve both student health and achievement. A back-of-the-envelope analysis suggests that for each effect, benefits were far in excess of costs. The average retrofit required only $8,110 in our sample, suggesting diesel engine retrofits can be at least three times more cost-effective than class-size reductions for achieving a given test score improvement.

2 Background

School bus diesel emissions are a public health concern because school buses are ubiquitous, concentrated in residential areas, and dirtier than most vehicles. Monahan (2006) finds that California

school buses were nearly twice as polluting as the average tractor-trailer. Such a surprising discrepancy is due to the age of the bus fleet; a 30-year-old school bus can produce two or three times as much on-board pollution as a 3-year-old bus.[5] The discrepancy also arises because diesel engines are dirtier than gasoline engines, contributing to a third of nitrogen oxide emissions and a quarter of particulate matter emissions despite being a smaller fraction of the automobile fleet.[6] School buses contribute to pollution exposure both for individuals spending more time near bus stops and along bus routes, but they are highest for passengers of the vehicle.[7] In fact, Zuurbier et al. (2010) find that riders of diesel buses had twice as much exposure to air pollution as carpoolers.

2.1 Emissions and Health

Exposure to air pollution worsens infant and childhood health. Diesel emissions contain smoke-related particulate matter, nitrogen dioxide, gaseous aldehydes, carbon monoxide, and toxic polycyclic hydrocarbons. The latter are potent carcinogenic compounds that are more stable when they diffuse into airborne water vapor, allowing them to reach deep into the lungs when inhaled.[8] For this reason, diesel exhaust may cause immediate short-term adverse pulmonary effects by decreasing the membrane potential of epithelial cells in the lungs.[9] There are also longer-term effects of diesel exhaust exposure; one cohort study of urban bus drivers in Denmark found that just three months of bus driving was associated with an increased risk of six types of organ-based cancers and all malignant tumors.[10] Young individuals are especially vulnerable to this form of pollution. Worse air quality has been linked to child lung function growth disparities of 3 to 5 percent, or four times the effect of second-hand cigarette smoke, in more-polluted areas, while exposure to in-traffic air pollution is associated with lower lung capacity, lower forced expiratory flow, and asthma development.[11] Two recent studies exploit variation in bus pollution at the census block level in New York City. The first (Ngo, 2015) found that increasing emission standards over time reduced emergency department visits for respiratory diseases among residents living within a few hundred feet of a

[5] Harder (2005).
[6] EPA (2003).
[7] Marshall and Behrentz (2005); Xu et al. (2016).
[8] Commins et al. (1957); Muzyka et al. (1998); Waller et al. (1985).
[9] Stevens et al. (2010).
[10] Soll-Johanning et al. (1998).
[11] Beatty and Shimshack (2014); Clougherty and Kubzansky (2008); Gauderman et al. (2005); Gendron-Carrier et al. (2018).

bus route. A second (Ngo, 2017) exploits variation in bus age, and thereby pollution levels, finding that children born to mothers who lived close to bus routes with older (dirtier) buses saw modest reductions in infant birth weight and gestational age compared with those living near routes with newer, cleaner, buses.

2.2 Emissions and Academic Performance

Past work has identified three mechanisms through which air pollution may impact test scores: attendance changes due to pollution-related illness, short-term disruptions in attention and cognitive performance, and long-term negative influence of pollution exposure on brain development. Currie et al. (2009) demonstrate that higher pollution levels over six-week periods are associated with more student absences, which may indirectly impact student learning. More directly, ultrafine particles in air pollution, particularly in diesel emissions, deposit in the prefrontal cortical and subcortical regions of the brain via the olfactory bulb, leading to heightened inflammatory response, white matter lesions, and behavioral and cognitive impairment.[12] Such cognitive impairment is observable in standardized test scores, and the negative effects stem from both contemporaneous and long-term exposure.[13]

2.3 Emission Reduction Programs

The well-known dangers of pollution from school bus diesel emissions led the United States Congress to spend $200 million per year from 2007-2012 to retrofit buses under the Diesel Emissions Reductions Act. Separately, the Clean School Bus Grant Program spent $110 million in 2005 and 2006. These grants pay for any one of four types of engine retrofits in our sample: diesel particulate filter, diesel oxidation catalyst (DOC), flow-through filter, or a closed crankcase filter (i.e. closed crankcase ventilation system (CCV)). Since the average diesel particulate filter costs between $5,000 and $10,000, engine retrofits have the potential to be a cost-effective means of reducing ambient air pollution and the health concerns associated with them.

The most common type of retrofit, a diesel particulate filter, can decrease overall emissions of particulate matter (PM) between 60 and 90%.[14] The effect of these filters on PM levels inside

[12] Calderón-Garcidueñas et al. (2012); Freire et al. (2010); Guxens and Sunyer (2012); Sunyer et al. (2015).
[13] Chen et al. (2017); Ebenstein et al. (2016); Ham et al. (2014).
[14] Biswas et al. (2009); EPA (2003).

the bus cabin is more modest at between 15-26%.[15] Emissions reductions of heavy metals from a diesel particulate filter are more substantial, in the range of 85-95%.[16] Emissions of other harmful compounds, such as total hydrocarbons and carbon monoxide, can be reduced to background pollution levels.[17] Finally, reductions of nitrogen oxide emissions can be significant; Tate et al. (2017) found that retrofitting the bus fleet in York, UK, would reduce city-wide levels of nitrogen oxides by 6-7%. These benefits appear to be fairly persistent with good engine maintenance and the use of low-sulfur fuels. Another study found that the reductions in PM of 95% by mass remained after four years of road exposure.[18] Taken together, the existing scientific evidence suggests that retrofits dramatically reduce students' exposure to potentially harmful compounds.

Our work builds most directly on Beatty and Shimshack (2011), who use a series of roughly 4,000 school bus retrofits in Washington state between 1996 and 2006. They match retrofit data and hospital admissions at the district-month level. The authors find that districts with retrofits saw significant and sizable reductions in asthma and pneumonia-related visits for both children and adults, with estimated benefits of nearly 7 to 16 times the cost of retrofit investments. In a related article that focuses on direct measures of exposure to pollution, Adar et al. (2015) measure pollution and health of 275 elementary school bus riders in Seattle and Tacoma, Washington, during a retrofit program from 2005-2009. The authors separately estimate the effect of four different emissions reduction progams – DOCs, CCVs, and fuel switching to ultra-low sulfur diesel (ULSD) or biofuels – on pollution exposure, health measures, and school absenteeism. They find significant effects of DOCs, CCVs, and ULSD use on on-board particulate levels. Health benefits (increased lung functioning measures) were found from DOCs and CCVs only for students with persistent asthma.

We build on this prior literature in several ways. First, we have different measures of student health: aerobic capacity and BMI from FitnessGram tests. Beatty and Shimshack (2011) use hospital visits, and Adar et al. (2015) use measures of lung functioning.[19] Our health outcome measures

[15] Hammond et al. (2007).

[16] Hu et al. (2009).

[17] Jiang et al. (2018). Note that Zhang and Zhu (2011) find that retrofits significantly decreased tailpipe emissions but had no significant effect on on-bus ambient air quality, while Li et al. (2015) show that tailpipe emissions do in fact enter the cabin. Borak and Sirianni (2007) conduct a meta-analysis and concludes that control technologies like retrofits can in fact eliminate "self-pollution" from diesel exhaust into bus cabins.

[18] Barone et al. (2010).

[19] Adar et al. (2015) use forced expiratory volume in one second (FEV1) and forced vital capacity (FVC) as measures of lung functioning. These measures are useful figures for diagnosing lung diseases such as COPD or emphysema, but

are likely to better capture the effect of diesel emissions on student health because VO_2 max conveys general cardiovascular health rather than lung function, therefore representing the observable consequence of lower lung functioning. Our outcome also captures the health of all students instead of merely those visiting a clinic for acute lung conditions, thereby capturing the effect on the average student instead of only those likely to visit a clinic. Second, we provide potential placebo measures using a non-respiratory health outcome, BMI. Third, ours is the first study we know of to examine the effect of retrofits on academic performance, allowing us to tie together two largely separate literatures on health and academic performance.

2.4 Retrofits in Georgia

The Georgia retrofit program started as the Adopt-a-School Bus program in 2003, a collaboration between the state Environmental Protection Division, school districts, and businesses to improve the well-being of students. The project's goals were to implement any of four emission reduction retrofit devices, reduce bus idling, and increase use of ultra-low sulfur diesel.[20] The project has since been funded by a wide variety of sources and grants. The EPA Clean School Bus grant program provided three separate grants in 2004, 2005, and 2006. The EPA's Diesel Emissions Reduction Act (DERA) also sponsored two retrofit grant cycles in 2009 and 2014 that collectively paid for 182 school bus retrofits. The US Department of Transportation sponsored the program under its Congestion Mitigation and Air Quality Improvement (CMAQ) Program, which contributed $11.2M to retrofit 1,890 buses. The staggered funding and implementation lags allow us not only to compare retrofitting and non-retrofitting districts, but also to exploit the timing of retrofits among retrofitting districts to secure causal identification.

Over the relevant sample period from 2007-2017, 2,656 buses were retrofitted with at least one type of modification. 1,160 of these bus retrofits involved a diesel particulate filter, 1,394 added a diesel oxidation catalyst, 58 installed a flow-through filter, 244 added a closed crankcase filter, and 188 buses were replaced early. We do not observe any information on the use of ULSD fuel, but we know from communication with the Environmental Protection Division that retrofit grants

they do not measure cardiorespiratory fitness *per se*. The FitnessGram aerobic capacity test we employ is designed to capture VO_2 max, the maximal oxygen uptake at peak performance. VO_2 max is used as a broader indicator of health (see Ross et al. (2016)).

[20] Idling reductions were a statewide effort.

stipulated the use of ULSD fuel to preserve the new engine parts. Moreover, EPA diesel fuel standards required the use of ULSD on all vehicles starting in 2010.

3 Data

Our data come from four sources, providing information on health, achievement, retrofits, and the Georgia bus fleet in general. Since we observe school bus retrofits at the district level, we aggregate data to that unit of analysis. We describe each data source, advantages, and limitations in turn below.

3.1 Health

Our first data source contains health information from the Cooper Institute's FitnessGram examination. The FitnessGram examination is a series of mandatory tests administered annually to all Georgia public school students who are in a physical education class. Many other states use FitnessGram as well, and the results of the FitnessGram tests are used widely in studies on student health.[21] According to the Georgia Department of Education's 2016 Fitness Assessment Program Report, 1.1 million students in Georgia (74%) participate in the examination. Since physical education requirements differ by age, the participation rate for elementary school students is 94%, while for middle school and high school students the rate is 71% and 49% respectively. Since our study covers several years, most students should be included at some time in the sample window.

Several tests are involved in a FitnessGram examination, including tests of aerobic capacity, body mass index, curl-ups, push-ups, and sit and reach (a measure of flexibility).[22] We limit our analysis to tests for aerobic capacity – a measure of cardiovascular fitness likely to be affected by exposure to diesel pollution – and for BMI – a potential placebo.[23]

[21] Anderson et al. (2018); Castelli et al. (2007); Edwards et al. (2011); Fahlman et al. (2006); Murray et al. (2012); Welk et al. (2010).

[22] Records of these assessments are kept by the Georgia Department of Education Physical Fitness Division, which annually reports school-level results separately for male and female students. For each school-gender-test combination, measures include the total number of attempts, the average performance, and the percentage of students attaining "healthy fitness zone" (HFZ) status. Depending on whether the aerobic capacity or BMI is higher than a benchmark figure determined for each student's age, weight, and gender combination, a student may be assigned to healthy fitness zone status.

[23] We exclude curl-ups, push-ups, and sit and reach from our analysis because they are not completed by a large proportion of the student body.

Aerobic capacity is the maximum rate at which oxygen can be taken up and utilized by the body during exercise. It is measured by FitnessGram through an exercise called the PACER (Progressive Aerobic Cardiovascular Endurance Run) test, also called a multi-stage fitness test, a "beep test", or a shuttle run.[24] Physical education instructors administer the test and record results according to instructions provided by the Cooper Institute. The school-level average VO_2 max, as computed from either the student-level number of laps completed on the PACER test or the timed performance on a one-mile run, is our observed outcome measure.[25] The FitnessGram assessment also directly measures each student's BMI, which is defined as a student's mass in kilograms divided by her height in meters squared. The CDC defines healthy and unhealthy levels of BMI for children based on their percentile rank among all children of a given age and sex.

The first and second years of FitnessGram aerobic capacity information collected by the state, 2011-12 and 2012-13, are not consistent with the remaining years.[26] These early years feature many average VO_2 max values that are simply not observed in later years. More troublingly, some of these very low average VO_2 values correspond to unrealistically high levels of healthy fitness zone attainment. The indiscrepancies may result from a few possible factors, although we cannot diagnose the precise origin of the issue.[27] Since the unreliability of the data is primarily a concern for the roll-out year of 2011-12 and much less so for 2012-13, while the different FitnessGram version is

[24] In the test each time students hear a timed electronic beep they have a set amount of time to run 20 meters (from one line to the other). The exercise ends for a student the second time she cannot finish the 20 meters within the set amount of time. At the end of each minute students hear 3 beeps letting them know that the amount of time they will have to finish the 20 meters has been reduced. A student's score is the number of laps she completed before her second failure to complete the 20 meters within the allotted time. Some schools actually use a one-mile run test to assess aerobic capacity. We do not observe the test employed, however both tests are converted to a comparable scale of VO_2 max. See Boiarskaia et al. 2011 for additional information on how these two tests are converted to the same measurement of VO_2 max, and Blasingame (2012), which finds that both assessment types accurately capture VO_2 max and are consistent with each other.

[25] Given age and weight, the number of laps completed by a student can be used to determine the student's maximal aerobic capacity, or VO_2 max. The Cooper Institute approximates this value based on a functional transformation of the number of laps completed and the student's age. For more information, see the Cooper Institute FitnessGram Reference Guide.

[26] Figure A1 displays the extent of inexplicable values in 2011-12 and 2012-13, showing how the otherwise tight linear relationship between percent of students in the healthy fitness zone and the average VO_2 max, which we see in the 2014-2017 data, is dramatically less reliable in the first two years.

[27] One potential cause is that schools calculated VO_2 max using the FitnessGram version 8 equation in 2011-12 and 2012-13, whereas in later years they use the conversion equation from FitnessGram version 9. Second, roughly one third of schools implement a one-mile run test while the remaining schools use the PACER test. Although both have been converted to units of VO_2 Max in our data, the correlation between VO_2 max and performance on the one-mile walk is slightly lower. See Blasingame (2012) for a thorough treatment of these two issues. The study finds that the one-mile run is less correlated to actual VO_2 max than the PACER test (correlation coefficients of .84 and .93), but both assessment types and estimation equations are consistent and generally accurate. Third, coaches may have half counted PACER laps, effectively counting a "down and back" as one lap rather than two. We suspect this issue because more-recent official coaching instructions specifically advise against this counting practice.

common to 2011-12 and 2012-13 and the one-mile run test is used by some schools across the entire sample window, it seems likely that the unreliable observations in 2011-12 are primarily an issue of accidental half-counting by coaches administering the test for the first time. This is consistent with the findings of Blasingame (2012) that differences between one-mile run and the PACER test and between FitnessGram versions 8 and 9 are minimal.

To account for this issue while preserving as much data as possible, we take a rule-based approach to identifying schools that most likely have contaminated scores, dropping any school-level observations below the minimum score by gender that we observe across all years in which we are confident of the data (those after the 2012-13 school year). In section 5.3.2 we explore the robustness of our results to a wide variety of alternative methods for dealing with this issue, including dropping the 2011-12 school year entirely, confirming that our main results are indeed quite conservative.

The first panel of Table 1 presents summary statistics of the FitnessGram tests for aerobic capacity (AC) and body mass index (BMI) aggregated to the district level. We take the district average as our outcome measure because treatment in our data is at the district level. Average values were converted from school- to district-level by calculating the sum of weighted school averages for each district, where the weight is the proportion of a district's attempts taken at that school.[28] The attempts divided by enrollment is an approximation of the proportion of students completing a FitnessGram examination in each district. AC and BMI were the two most common FitnessGram examinations, though less than half of students in a district completed the AC exam, while about two-thirds of students completed a BMI examination in any given year.[29]

[28]For example, district i's average aerobic capacity in a given year is $y_{it} = \sum_{s=1}^{N} x_{st} \frac{a_{st}}{a_{it}}$ where x_{st} is the school average in year t and a is the total attempts on the relevant FitnessGram examination for each school s in district i and year t. Alternatively, the weights could be school-level and district-level enrollment instead of total attempts, but this aggregation procedure overemphasizes schools that have lower levels of FitnessGram participation, such as high schools.

[29]Some students are not tested because children below 3rd grade do not take the test, and any students who are not in a physical education class also do not take the test. Additionally, tests administered to fewer than 25 students in a school are coded as zeros to protect privacy, hence some school observations are missing. In Appendix Table A3, we find no relationship between FitnessGram attempts on aerobic capacity and bus retrofits. It is also impossible to know whether the total attempts reported by the state reflect multiple attempts by the same student. This could introduce noise if, for example, districts compensate for lower performance by allowing their students more attempts, which would tend to mute physical fitness differences across districts. We also test for this possibility in Appendix Table A3.

3.2 Academic Achievement

Our second source of data includes information on student test scores, enrollment levels, and demographics from the Georgia Department of Education (GADOE), which provides school-level data from 2006-07 to 2016-17. Only English language arts (ELA) and math end-of-grade 3rd-grade through 8th-grade test scores are reported throughout the sample window, so we focus on these exams. The state's recorded information includes the average raw scale score of students in each grade and the number of student test takers for each test. We normalize scale scores using the state mean and student-level standard deviation, and then average over grades and schools using weights for the number of test-takers. This yields a district-level average performance, in terms of student-level z-scores, for ELA and math in each year of the sample. From 2013-14 to 2014-15, the state changed its assessment regime from the Criterion-Referenced Competency Test (CRCT) to the Georgia Milestones Assessment System, with an accompanying change in scale and difficulty on the math end-of-grade exam. This is accounted for by normalizing within grade-year and including year fixed effects in our regression models.[30] The second and third panels of Table 1 display district-level schooling outcomes and demographic characteristics. Test scores are slightly higher for retrofitting districts,[31] though this may be confounded by the effects of the retrofits themselves. Attendance rates are virtually identical across retrofitting and non-retrofitting districts. On average, non-retrofitting districts are smaller, but have otherwise similar student compositions.

3.3 Bus retrofits

The third data source contains information on all bus retrofits from 2003-2018 and was provided through an open records request by the Georgia Environmental Protection Division (EPD). These data describe the type of retrofit performed in each district, the number of buses affected, the month and year of implementation, and the specific grant used to finance the retrofit. We use district-specific invoices for reimbursement for installation of retrofits to calculate the amount each

[30] Later, in Table A4, we drop the Milestones years from the sample. Aside from being a slightly different examination, there were widespread issues with the new computer-based assessment. The state notably decided not to use the Milestones examination for accountability purposes in 2015 and 2016.

[31] Standardized test score averages are different from zero because there are many low-performing districts with small student populations and a few high-performing districts with many students.

district paid for their retrofits.[32] Figure 1 maps retrofitting districts. The fourth panel of Table 1 shows that a typical retrofitting district improved 66 buses, or close to 19% of the bus fleet, in each retrofit cycle.

3.4 Bus manifest

We augment this with the Georgia Transportation Authority's manifest of all state school buses from 2010-2016. Since the bus manifest covers fewer years than for which there exist retrofits, information for 2007-2010 and 2017 is replaced with the value of the nearest available year in the sample.[33] The manifest includes specific bus identifiers, type of bus, capacity, and bus manufacturing details like make, model and year, fuel source, passengers, daily miles, and the number of students living within 1.5 miles of the school who are eligible to be riders. Some of these statistics are summarized in the last panel of Table 1. The variety of information provided by the bus manifest allows the creation of variables for the district-wide average student minutes spent in the bus, the district-wide bus ridership rate, and the proportion of district buses retrofitted for each grant. These three variables comprise our treatment measures. In our sample, the average bus rider spends a little less than 45 minutes on the bus each day. The average district has a 62% bus ridership.

4 Empirical Strategy

Our identification strategy exploits variation in the timing and location of retrofits across Georgia. We adopt a first-differences estimation strategy, which differences out any unobserved, time-invariant district attributes that might be correlated with retrofits and health or achievement. The estimating equation is as follows:

$$\Delta y_{it} = \beta R_{it} + \Delta X_{it}\gamma + \tau_t + \Delta\epsilon_{it}. \tag{1}$$

[32] Although we do not observe actual emissions pre- or post-retrofit, the EPD does provide predictions of the yearly and lifetime reductions of four pollutants (fine particulate matter (PM2.5), volatile hydrocarbons, carbon monoxide, and nitrogen oxides) using the EPA Diesel Emissions Quantifier. Because these are predicted emissions changes based on engineering models rather than measured or observed values, we do not use these data.

[33] Inclusion or exclusion of these years does not affect the sign or diminish the magnitude of the results, as we show in Appendix Table A5.

All variables are aggregated to the district (i) year (t) level as described above. Δ indicates a one-period change in a variable, e.g. $\Delta y_{it} = y_{it} - y_{it-1}$. The dependent variable y_{it} can be either one of the two health outcomes (aerobic capacity and body mass index) or one of the three schooling outcomes (math and English scores and attendance). Since many retrofitting districts experience more than one retrofitting episode, the model captures these year-on-year changes in health and schooling as a result of proportional changes in the share of buses retrofitted.

Our treatment variable, measuring district retrofits that occurred between time $t-1$ and t, is R_{it} (one can think of R_{it} as the change in cumulative retrofits between $t-1$ and t.)[34] We consider three different ways of measuring treatment intensity, R_{it}. The first measure is the proportion of the bus fleet retrofitted that year, termed *Percent Retrofitted*. For example, if a district retrofits 10% of its buses between $t-1$ and t, then $R_{it} = 0.1$. In this case, the magnitude of the coefficient on R_{it} shows the effect of retrofitting an entire fleet – going from all dirty buses to all clean buses.[35] The second measure is the proportion of the bus fleet retrofitted multiplied by the time-constant proportion of students in the district who are bus riders, termed *Percent Retrofitted * Ridership*. For example, if 10% of buses were retrofitted between time $t-1$ and t, and time-constant average bus ridership in district i is 50% of students, then $R_{it} = 0.05$. Here, the coefficient on R_{it} shows the effect of retrofitting an entire fleet in a district where all students ride the bus. This accounts for the fact that the impact of retrofitting should have a larger effect in districts where a higher fraction of students ride the bus. We use time-constant district averages for the proportion of students who are bus riders to avoid identifying changes off potentially endogenous ridership changes. Our third measure is the proportion of the bus fleet retrofitted times the fraction of students who are bus riders times the time-constant average duration of each bus ride in minutes per day. This is termed *Percent Retrofitted * Ridership * Trip Duration*. Here again we use the time-constant district average for bus ride minutes to avoid identifying effects off potentially endogenous changes in trip duration. Given two district-years with an equal proportion of buses retrofitted and an equal share of students who ride the bus, if one district buses students twice as far as the other, we should expect larger effects in that district.

[34] In other words, we could also have modeled this as ΔR_{it}^{cumul}, the change in cumulative retrofits. This causes difficulties when we interact R with the share of students who are bus riders because we do not want to identify variation resulting from potentially endogenous changes in ridership.

[35] The average proportion of the fleet retrofitted for the observed retrofits is 0.189.

Equation 1 includes the vector ΔX_{it}, measuring annual changes in the following district-level student characteristics: percent of the student body that is Asian, Hispanic, African-American, male, English-language learner, eligible for free- and reduced-price lunch, or possessing of a disability. The vector ΔX_{it} also includes the following district-level changes in bus fleet characteristics: average bus age, to account for new buses replacing older models, the share of buses that are older models made before recent emissions regulations, de-meaned student ridership, de-meaned trip duration, and the share of buses that run on liquid natural gas, regular gasoline, and butane. We find little impact from their inclusion. τ_t is a schoolyear fixed effect.

Our identifying source of variation is the timing and magnitude of the retrofits. Differences in the share of students riding the bus and the average length of ride among riders add additional variation. An identifying assumption is that this timing is uncorrelated with any potential confounders that would affect health or academic performance. This assumption would be violated if, for example, retrofit timing was a function of expected changes in health or academic performance. Such endogeneity is unlikely in practice because funding allocation decisions were made by a state agency, the Environmental Protection Division, independently of any school district prerogatives. Moreover, the timing of bus retrofit completion varied greatly within grant cycles and across districts. Still, if this were true, we would also see changes in BMI as a proximate health outcome, which we test for. We might also be concerned with endogenous responses on the part of students and families through, for example, increased ridership in response to cleaner retrofitted buses. We employ several robustness tests to allay each of these concerns and discuss each in turn directly following our main results.

One could also estimate the model using district fixed effects. This, though, requires stronger assumptions than the first differences model, some that we likely do not satisfy. For example, the first differences model best captures immediate year-on-year changes, given that a large number of districts have multiple retrofit cycles. More importantly, we worry about serial correlation. First differences requires only that R_{it} is uncorrelated with $\Delta\epsilon_{it} = \epsilon_{it} - \epsilon_{it-1}$ where fixed effects requires R_{it} to be uncorrelated with $\epsilon_{it} - \overline{\epsilon_i}$ (i.e. that all errors are uncorrelated as opposed to uncorrelated changes in errors, which is a weaker assumption). In the absence of serial correlation, the fixed effects estimator has consistency advantages over first-differences, but as we show in robustness checks, Durbin-Watson statistics suggest that we do not satisfy this requirement, in particular for

academic outcomes which are highly serially correlated. Moreover, we have a relatively large number of time periods (10) compared to the number of individual observations (180), which again leads to advantages in the first differences model as fixed effects assumes $N \to \infty$ with fixed T. Regardless, as part of our many robustness tests we also report estimates from fixed effects regressions. We show that while point estimates are similar in nearly all cases, standard errors are larger under the fixed effects model, which is consistent with our concerns.

5 Results

5.1 Health

We present our main regression results for aerobic capacity and our placebo outcome, BMI, across all three measures of treatment R_{it} in Table 2. These regressions are based on Equation 1 and use data from 2012-2017. The first three columns present effects on aerobic capacity (AC), where the units represent VO_2 max, which is measured in milliliters of oxygen intake per kilogram minute. The second three columns present the effects on BMI. The coefficient in column 1 implies that if a district retrofitted 100% of its fleet, average VO_2 max would increase by 1.8 units, or about a 4% increase relative to the baseline mean of 41.16. Since the average retrofit affected 19% of the bus fleet, the average retrofit improved district-wide aerobic capacity by 0.33 milliliters of oxygen per kilogram minute.

Columns 2 and 3 use the alternate measures of the treatment effect R_{it}. In column 2 it is the percent of buses retrofitted multiplied by the percent of students who ride the bus. This coefficient implies that if a district had 100 percent ridership *and* retrofitted its entire bus fleet, average student aerobic capacity would increase by 2.4 units, or about 6 percent of the mean. The average bus ridership rate is 62%, so this implies that the average retrofit (19% of the fleet) in the average district increases aerobic capacity by 0.28 milliliters of oxygen per kilogram minute. Finally, column 3 sets R_{it} to the percent of the bus fleet retrofitted times the ridership rate times the average trip duration. The coefficient implies that, if all buses in a district are retrofitted and all students ride the bus, then each additional minute of bus riding for students in this district is associated with roughly 0.041 units increase in VO_2 max. Since the average trip duration is 46 minutes, this implies

that the average retrofit in the average district increased aerobic capacity by 0.21 units VO_2 max.[36] Thus our point estimates, when scaled, are roughly consistent across specifications in the range of 0.2 to 0.4 units VO_2 max. Given that there is little variation across retrofitting districts in the ridership share and trip length, we do not find this result surprising.

We next turn to our placebo health outcome, BMI. In the final three columns of Table 2 we find that estimates are effectively zero in all cases. Although directions suggest lower BMI, the coefficient on our main estimate (-0.24) is equal to approximately 1% of BMI. We take this as suggestive evidence that retrofits were uncorrelated with general health trends across treatment and control districts.

We next break out results by gender and school level. Table 3 displays male and female aerobic capacity results in the full sample across elementary, middle, and high schools. These results reveal two pieces of information. First, estimates are comparable for male and female students. While point estimates are different across gender for elementary school students, the coefficients for male and female students at a given level are not statistically different from one another. Second, effects are highest among elementary school students. We find noisy and in fact negative effects for boys in middle school. Although we are unable to explain this, we believe it relates to influence of outliers in the middle-school assessments and the likely re-assessment of physical education classes to selectively lower-quality students after one mandatory year of the course. The consistency across elementary male and female estimates contradicts the hypothesis that differential incidence of childhood asthma in young boys would exert some influence on these relative effect sizes (Bjornson and Mitchell, 2000). As has been shown in other work (Beatty and Shimshack, 2011), children with asthma are more susceptible to the negative effects of air pollution.

5.2 Academic Achievement

We present our main regression results on three academic outcomes in Table 4. These regressions include years 2007-2017, since we observe test scores for more years than we observe FitnessGram outcome measures. In columns 1-3, the dependent variable is a z-score of average English (ELA) test scores, normalized to the student-level standard deviation, for grades 3-8. The coefficient in column 1 implies that retrofitting an entire fleet would raise ELA scores by 0.09 standard deviations.

[36] $= 0.189 * 0.62 * 46 * 0.041$

This represents an achievement differential slightly larger than that observed between students of a rookie teacher and those of a teacher with five years of experience.[37] The average retrofitting district retrofitted 19% of the fleet, suggesting an average increase in ELA scores of 0.017 standard deviations per retrofit cycle. In column 2, the treatment effect R_{it} is the share of buses retrofitted times the share of students who ride the bus. The point estimate suggests that retrofitting an entire bus fleet with 100% ridership would increase student test scores by 0.143 standard deviations. The average retrofit (19% of the bus fleet) for the average district (61% ridership) increased scores by 0.017 standard deviations according to this point estimate, which is identical to the result in column 1. Column 3 shows that each minute of bus riding in a 100%-retrofitting district with 100% ridership is associated with a 0.003 standard deviation increase in ELA scores. Based on this, the average district's retrofit increases ELA scores by 0.016 standard deviations, which is consistent from specifications (1) and (2).

The results on math test scores (columns 4-6) are also positive but only about one-half as large as the ELA results and not statistically distinguishable from zero. This is consistent with Ham et al. (2014) who find that particulate matter, and especially PM2.5, tends to affect ELA scores more than math scores. Specifically, they find that PM2.5 lowers math scores by 60% less than ELA scores, which is similar to our findings. The last three columns of Table 4 show that there is no effect of retrofits on average attendance rates. Since the mean attendance rate is 0.95, there is little margin for gain. This contrasts with the negative attendance effects found in Adar et al. (2015). In Table 5, we show how the percentage of a bus fleet retrofitted affects ELA and math z-scores among elementary and middle school students.[38] Consistent with the health estimates, effects are larger in elementary schools than in middle schools. For both elementary and middle schools, the effects on math are positive but indistinguishable from zero.[39]

[37] Rice (2010).

[38] We do not have test scores by gender, nor do we have them for high school students.

[39] In Appendix Table A1 we display results dis-aggregated by grade. The grade-level performances are consistently in the same direction as the main academic estimates, and achieve significance in at least one grade for each ELA and math test scores. Interestingly, grade-level effects suggest larger impacts for students more likely to sit at the back of the bus– those in 4th, 5th, and 8th grade– which is consistent with bus self-pollution from diesel exhaust.

5.2.1 Results by Retrofit Type

In Table 6, we present results by type of retrofit for each of our academic and health outcomes. There were few episodes of closed-crankcase filter retrofits (244), and fewer of flow-through filter retrofits (58). In fact, there were none of these retrofits over the sample period during which we observe aerobic capacity and BMI records from the FitnessGram examination. Nevertheless, diesel particulate filters (1,160 retrofits, or 44%) and diesel oxidation catalysts (1,394 retrofits, or 43%) had a positive and roughly consistent effect on both ELA and math test scores. Adar et al. (2015) found that implementing DOCs and CCFs both had an effect on attendance, with larger and more significant effects for DOCs. This is consistent with our findings for DOCs only, the discrepancy likely caused by the low number of CCF retrofits. We add to Adar et al. (2015)'s findings by testing for effects on diesel particulate filters, which appear to have a larger effect on ELA, math, and aerobic capacity. Since DPFs are expected to eliminate 60-90% of fine particulate matter, while DOCs eliminate 10-50% of fine particulate matter in Adar et al. (2015), this finding appears reasonable.

5.3 Robustness and Alternate Specifications

5.3.1 Academic Achievement Pre-Trends in Retrofitting Districts

One might be concerned that retrofitting districts have different pre-treatment trends that drive the results. This possibility is difficult to test directly because there is no uniform year of treatment across retrofitting districts. For this reason, there are no uniform pre-treatment or post-treatment years. Many retrofitting districts also had multiple retrofit cycles. To assess the possibility of differential pre-trends, we therefore plot academic achievement outcomes from 2006-07 to 2011-12 across retrofitting and non-retrofitting districts in Figure 2. We plot results before 2013 because this was the modal retrofit year with nine retrofits. We note that 25 retrofits occur before this year, so we may expect the slope trends to increasingly differ by the extent to which the retrofits impact academic outcomes. Nevertheless, the trends appear close to parallel over this period. We do not plot pre-trends for our health outcomes because of the shorter window over which we observe these outcomes and the notable issues with aerobic capacity information in the roll-out year of the program (as discussed in section 5.3.2).

5.3.2 Aerobic Capacity Data

As discussed earlier, the early FitnessGram results contain inconsistencies, so we apply a rule-based approach in which we eliminate implausible values. In Appendix Table A2 we re-estimate our main specification, using the share of buses retrofitted, across different cutoff values to demonstrate how our results vary across different rules of thumb. The first five columns of the table show results for cutoffs set at $15, 20, 25, 30, 35$. These represent dropping school-level aerobic capacity results below the given value in 2011-12 and 2012-13 (although, in practice, almost all removed values are in 2011-12). In column 6 we show our preferred cutoff of 26 for females and 30 for males for reference, the lowest observed values after 2012-13. In column 7 we apply an alternate rule where we eliminate schools for which we observe a jump of more than 6 in Aerobic Capacity – equivalent to 15 percent of the mean – between 2011-12 and 2012-13 as an indicator of reporting issues in the first year. In column 8 we show the full data, not dropping any schools, and in column 9 we show effects if we drop school year 2011-12 entirely. With the exception of the specification in columns 7 and 9, results are similar in magnitude across specifications. Eliminating problematically low observations affects the standard errors, as we would expect. In column 7, when we drop implausibly large jumps, estimates double, and when we drop the first year of data entirely in column 9, effect sizes increase over four-fold, from 1.8 to 7.1. While we are more confident in these estimates, we take the conservative case of only dropping problematic observations as our preferred estimate.

5.3.3 Correlation of Proportion of a Bus Fleet Retrofitted with District Characteristics 2007-2017

We address the potential for retrofits to affect participation in the FitnessGram test, possibly due to increased health status, in the first panel of Appendix Table A3. In columns 1 and 2, we regress the participation rates for aerobic capacity and BMI FitnessGram tests, measured as the total number of test attempts divided by the district enrollment, on the percent of a bus fleet retrofitted. We find no discernible relationship between district retrofits and the share of student who are tested in aerobic capacity. If anything the point estimate suggests a small negative relationship. We find a similar pattern for BMI tests, suggesting that districts with more retrofits see a marginally higher rate of BMI testing, though again the estimate is noisy. In column 3 we test for changes in ridership,

potentially resulting from an increase in the share or number of students riding the bus as a result of reduced emissions. We find a reasonable precise null effect, suggesting that cleaner buses do not increase ridership. In the same table, we demonstrate the relationship between the proportion of a bus fleet retrofitted and changes in bus fleet characteristics, student demographics, and student characteristics. We observe a statistically significant relationship in only one case; the proportion of students with disabilities in a district is positively related to the proportion of a bus fleet retrofitted. We believe it is unlikely that retrofits would change disability status among students and take this as a spurious correlation. Moreover, we control for changes in the share of students with disabilities in all regressions, it is possible that this correlation may reflect unobserved changes in district health or achievement.

5.3.4 Milestones Test Sensitivity

The roll-out of a new Milestones exam (Georgia's end-of-year test) in 2015 resulted in a large decreases in math scores in several of Georgia's largest districts, many of which received retrofits. The decrease was caused by complications in the new internet-based math examination where several districts had computers "freeze," causing severe disruption to test-takers.[40] As a result, those exams were not used to calculate district performance for state requirements, student retention, or graduation.[41] Because retrofitting districts are primarily Georgia's larger districts, which were those who adapted to computer based tests, raises concerns that this could be a confounding factor in our test score analysis. When we drop the Milestones years 2015-2017 from the sample, in Appendix Table A4, the results are qualitatively similar to our main specification, although math scores are larger in magnitude and more precise. Since no districts retrofitted after 2015, this change is not correlated with contemporaneous treatment, but rather shows a decline in test score post-treatment for these districts.

5.3.5 Exclusion of Interpolated Bus Manifest Data

The district bus manifest covers 2009-10 to 2015-16. We fill in the remaining years by substituting the value of the nearest chronological neighbor for each year. For example, a district's 2016-17 value

[40] Cobb, Dekalb, Cherokee, and Gwinnett counties all suffered from these computer glitches.
[41] See this article and this article for more information.

for total buses is set equal to the number of buses it had in 2015-16. Linear interpolation was ruled out because it created unrealistic values for some districts with large changes in their bus fleet. As shown in Appendix Table A5, our results are unchanged by the exclusion of years for which we lack information on district bus fleets. In fact, excluding these years improves the precision of both our math and ELA point estimates.

5.3.6 Timing of Retrofit Treatment

There are two sources of imprecision with respect to the timing of treatment. First, the FitnessGram test may be in fall, spring, or both, while the end-of-grade tests are uniformly in April-May.[42] Second, the date of the bus retrofit reimbursement invoice, which we use as a proxy for the date of retrofit completion, imperfectly corresponds to the date when the buses are first used. If the timing of a retrofit comes before April of the year in question, the retrofit is counted as occurring in that school year even if some of the FitnessGram tests may have occurred before the retrofitted buses were active. This may affect the results of some FitnessGram tests while leaving the test score results unaffected. On the other hand, buses completed in a retrofit before April may not actually be used until the following school year due to implementation lags, which would mean our baseline treatment year assignment is too early to pick up changes in test scores. In Appendix Table A6 we show our baseline treatment assignment and explore a placebo timing treatment that assigns the year of the retrofit to one year in advance of the year of the retrofit completion invoice. These results, presented in the second panel of Table A6, demonstrate that the assigned treatment timing is not inconsequential, as no estimate is significant when adopting a placebo treatment year. In Panel III we assess the possibility that our treatment assignment for retrofits occurring after January is too early by assigning the same fiscal year to any retrofits completed before January and the subsequent fiscal year to any retrofits completed after January. Under this treatment year assignment rule, the results are the same for each outcome except for math test scores, which are now positive and significant. We take this as suggestive evidence that our baseline treatment assignment is not too late to capture changes in aerobic capacity, although it may be too early to pick up changes in academic achievement for some districts.

[42] Across the state we know that two-thirds of FitnessGram exams are given in Spring and one-third in Fall, although we do not know the breakdown by district.

5.3.7 Fixed Effects vs. First Differences

In Appendix Table A7 we re-estimate our main results now using a fixed effects specification. In the top panel we show our main results from the first differences model for reference. In Panel II we show estimates from a standard fixed effects model. This is:

$$y_{it} = \alpha + \beta R_{it} + X'_{it}\gamma + \tau_t + \phi_i + \varepsilon_{it} \tag{2}$$

Where ϕ_i is a district fixed effect. Point estimates for all outcomes are similar, though noisier, with the exception of math scores which are now zero. At the bottom of the table we show Durbin-Watson test statistics indicating a high degree of serial correlation in test scores, though less so for health measures. Given that the math estimates were zero in our main specifications, a smaller coefficient here does not change conclusions. For ELA, we find a marginally larger point estimate and substantially larger standard errors. Effects on aerobic capacity are similar though noisier as well. Taken together, Table A7 suggests efficiency gains from first differences as expected in the presence of serial correlation, and that our conclusions are not substantively altered by our modeling choice.

As a final check, we use the FE model to add an additional test for trends. To do so, we re-estimate Equation 2 above and include a lead of R_{it}, which is $R_{i,t+1}$, which is a test for pre-trends. In panel III of Table A7 we show results (the final year of data is dropped because there can be no lead). We show that academic results are similar to our main estimates, and importantly, we cannot reject that all leads are zero. We do find that the lead for aerobic capacity is large, but this is due to the first year of data.

6 Cost-Benefit and Cost Effectiveness Analyses

We conduct back-of-the-envelope calculations of the costs and the benefits of bus retrofits. We examine health benefits in terms of both reduced mortality and reduced cardiovascular disease, as well as benefits from increased test scores. We note that this does not account for spillover effects on non-treated members of the community who are exposed to lower pollution levels overall. Additionally, we compare the cost of achieving the education benefits from the retrofits to the costs

of achieving similar gains from class-size reduction to provide a cost effectiveness analysis.

6.1 Costs

The total amount awarded for district bus fleet retrofits in Georgia is $26 million. However, certain retrofits occurred before our sample window. Moreover, a large portion of funds went to purchasing new buses to replace older ones. We separate the amount awarded for bus replacement from the amount spent on retrofits using invoices detailing each district's reimbursement for completing their retrofit. These reimbursements include the cost of parts, labor, and daily usage of a repair bay. The total amount spent on engine retrofits is $12.6 million, with the average district spending $8,110 per retrofitted bus. The average district has 111 buses, so the cost of the average district retrofitting 10% of its fleet is $90,000. For comparison, the cost of one regular new bus is roughly $130,000, while a new hybrid or electric bus is $360,000. Replacing 10% of a fleet with new diesel or hybrid buses would therefore cost $1.4M - $4M, an order of magnitude greater than the cost of engine retrofits.

6.2 Benefits - Health

We focus on the health benefits in terms of increased aerobic capacity, which is the most persistent result. Our preferred specification is column 1 of Table 2, which indicates that a ten-percentage-point increase in the percentage of buses retrofitted is correlated with a 0.18-unit increase in the measure of aerobic capacity. The units we observe for the aerobic capacity measure are milliliters oxygen per kilogram minute (mL/min/kg); these units have already been converted into a measure of VO_2 max from the number of PACER laps completed using a standard conversion factor provided by the FitnessGram test manufacturer. From this conversion we conclude that a ten-percentage-point increase in the percentage of buses retrofitted is correlated with a 0.18-unit increase in VO_2 max. We convert the VO_2 max effect measure from units of mL/min/kg to units of metabolic equivalent (MET) by dividing the VO_2 max in mL/min/kg by 3.5, yielding a change in MET of 0.05 for a retrofit of approximately 10 percent of a district's bus fleet.[43]

Several studies document and measure the benefits from increased aerobic capacity (or car-

[43] Castillo-Garzón et al. (2006).

diorespiratory fitness).[44] Kodama et al. (2009) conducts a meta-analysis and finds that a 1-MET higher level of VO_2 max is associated with a 13% decrease in the risk of all-cause mortality and a 15% decrease in the risk of cardiovascular disease (CVD).[45] However, this meta-analysis was conducted on studies of adults, not children. Other studies examine the effect of cardiorespiratory fitness on children's CVD outcomes[46], but do not provide an estimated magnitude of a causal effect from VO_2 max.

We thus use two different measures of the valuation of health benefits from aerobic capacity increases. First, we use the meta-analysis of mortality effects reported in Kodama et al. (2009) for adults and extend them to childhood mortality: a 1-MET increase in VO_2 max is associated with a 13% decrease in mortality risk. The baseline childhood mortality rate in Georgia among 5-12 year olds was 13.3 deaths per 100,000 population in 2016.[47] We use a standard value of a statistical life (VSL) of $7.4 million.[48] The average district in Georgia has about 9,000 students. Thus, if an average district's average MET unit of VO_2 max increased by 0.05 units (the effect size 1.8 scaled to represent a district retrofitting 10 percent of a its buses and divided by 3.5 to convert to MET units), the health valuation from reduced mortality for that district is $71.1.[49] Assuming a retrofit life of 10 years[50] and an annual discount rate of 3%, the present discounted value of the mortality reduction benefits is $624.69, a small fraction of the cost of retrofitting 10% of the bus fleet calculated earlier, $90,000. It is perhaps not surprising that the retrofits fail a cost-benefit analysis when the benefits are calculated only from reductions in mortality, since the baseline mortality rate for elementary-school-aged children is extremely low.

The second measure of the valuation of health benefits combines the result from Kodama et al. (2009) on the effect of aerobic capacity on cardiovascular disease (among adults) with results from Adamowicz et al. (2014) on the valuation of avoided CVD among children. Adamowicz et al.

[44] Several such studies are summarized in Institute of Medicine (2012), Chapter 5.

[45] Lakoski et al. (2015) finds also an association between aerobic capacity and adult cancer rates.

[46] Castro-Piñero et al. (2017); Ortega et al. (2008)

[47] https://oasis.state.ga.us/oasis/webquery/qryMortality.aspx#

[48] https://www.epa.gov/environmental-economics/mortality-risk-valuation#whatvalue

[49] The 0.05 MET increase = 0.00000665 PP decrease in the mortality rate = 0.00000960555 averted deaths per average district retrofit = $71.1 per district.

[50] Diesel particulate filters are often given a lifespan of 100,000 miles by the manufacturer, which represents 8 years with our sample's average yearly mileage of 12,960. However, DPF lifespan varies greatly depending on regular servicing and cleaning. Barone et al. (2010) show that DPFs are 95% as effective after four years, while Sappok et al. (2009) show that DPFs are half as effective at 188,000 miles, or roughly 14 years for the buses in our sample. We select 10 years to be consistent with prior work (Beatty and Shimshack (2011)), although the entire range (4-14 years) of possible lifespans lead to benefits far less than the costs of $90,000.

(2014) conduct a stated-preference survey of parents asking for their willingness-to-pay (WTP) for a reduction in the probability of their children being diagnosed with heart disease by age 75. They report a mean annual WTP to reduce that probability by one chance in one hundred of $5.62 for mothers and $4.08 for fathers; we use the mean of these two values ($4.85). Since this is an annual WTP, we interpret the total WTP for the one-in-one-hundred chance reduction in CVD to be the net present value of this annual WTP from age 11 until age 75, which equals $139.34.[51] Kodama et al. (2009) report a 1-MET increase in VO_2 max is associated with a 15% decrease in the risk of CVD. About one third of Americans have some form of CVD,[52] so a 15% decrease in the risk is equivalent to a decrease in the chance of 1 out of 20. Therefore, the benefit from a district retrofitting 10% of its buses is valued at $940,590 per district.[53] This is more than nine times greater than the cost of the retrofits. Because CVD is so prevalent (unlike childhood mortality), the valuation of even a modest reduction in its risk is quite high. These benefits do not take into account the value of lower pollution levels for non-students.

6.3 Benefits - Test Scores

Next, we calculate the benefit of the retrofits from a monetization of test score improvements. Chetty et al. (2011) estimate the effect of an increase in kindergarten test scores on adult earnings; they report that a one-percentile increase in test scores is associated with an increase of $94 in wage earnings at age 27 after controlling for parental characteristics. Assume that the wage benefit of $94 lasts throughout one's working years of age 25-54, and discount using an annual rate of 3%. Then, the one percentile increase in test scores is valued at $1,041.[54] The results presented in Table 4 indicate that retrofitting 10% of a district's fleet will increase the z-score of the ELA tests by 0.009 and of the math tests by 0.005. These improvements in z-scores are equivalent to percentile increases of 0.36 and 0.19, respectively. Using the average of these two values (0.275), and multiplying by the valuation implied by the Chetty et al. (2011) estimates, the benefit of retrofitting 10% of a district's fleet is valued at $2.57 million.[55] This is over 25 times greater than the costs of

[51] The survey sample in Adamowicz et al. (2014) includes just parents with at least one child aged 6-16 in the home, so we use 11 as the starting age.

[52] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5408160/

[53] The 0.05 MET increase = 0.0075% decrease in the probability of CVD = $104.51 benefit per child = $940,590 benefit for an average district with 9,000 children.

[54] $= \sum_{i=20}^{50} 94 * (1 - 0.03)^i$

[55] $= 0.275$ percentile points $*\$1,041$ per percentile point per student $*9,000$ students per district.

the retrofit.

Lastly, we compare the costs of achieving test score gains through bus retrofits to the costs of achieving those same gains through interventions studied in Chetty et al. (2011). The Tennessee STAR program reduced class sizes by seven students, which is expected to cost around $870 per student,[56] and it yielded a 4.81 average percentile improvement in test scores. Our estimates of the effects of the retrofits are that they yielded a 1.9 - 3.6 average percentile increase in test scores. The average school bus in our sample transports 66 students per day. Since the average cost per bus retrofit in our sample is $8,110, this translates to a cost of roughly $122 per student, or $34.1 - $64.7 per percentile point gain. The cost for an equivalent test score improvement is roughly three to six times higher for the STAR class size reduction than it is for the bus engine retrofits.[57]

7 Conclusion

We estimate the effect of retrofitting diesel school bus engines on student health and academic achievement in the state of Georgia. Retrofits have positive and significant effects on students' aerobic capacity, a measure of respiratory health, but no effect on body mass index, which we take as a placebo. Retrofits also have positive and significant effects on student English test scores, and a smaller and precise effect on math scores. Robustness checks reinforce our findings. Back-of-the-envelope calculations suggest that the benefits of the retrofits were much higher than their costs, and that the academic gains were achieved at a lower cost than they would have been through class size reductions.

This study could be extended several ways. First, use of individual-level data rather than district-level data may improve the precision of the results. Within-district variation in the exposure of students to the retrofits could be utilized if, for instance, individual student health records could be matched with bus routes. This could also allow for determining if treatment effects differ by demographic group. Second, data from other states could be analyzed to test whether the results from Georgia generalize elsewhere. Third, alternative health or academic outcomes could be examined. Linking students to other health outcomes, for example via Medicare data, may provide a valuable measure of health not picked up by FitnessGram scores. With a longer panel, long-term

[56] Reichardt (2000).

[57] The class size reduction cost $870 per student for 4.81 percentile gain = $181 per percentile point gain.

outcomes, including college attendance and labor market outcomes, could be examined. Fourth, we could test the effect of retrofits on outcomes other than health and academic performance such as non-cognitive skills.

Our results have plausible policy relevance. While bus retrofit programs are widespread, very little work has examined their effects. Policymakers interested in physical health and academic performance of children can use bus retrofits as another cost-effective policy tool.

References

Adamowicz, W., Dickie, M., Gerking, S., Veronesi, M., and Zinner, D. (2014). Household decision making and valuation of environmental health risks to parents and their children. *Journal of the Association of Environmental and Resource Economists*, 1(4):481–519.

Adar, S. D., D'Souza, J., Sheppard, L., Kaufman, J. D., Hallstrand, T. S., Davey, M. E., Sullivan, J. R., Jahnke, J., Koenig, J., Larson, T. V., et al. (2015). Adopting clean fuels and technologies on school buses. pollution and health impacts in children. *American journal of respiratory and critical care medicine*, 191(12):1413–1421.

Anderson, M. L., Gallagher, J., and Ritchie, E. R. (2018). School meal quality and academic performance. *Journal of Public Economics*, 168:81–93.

Barone, T. L., Storey, J. M., and Domingo, N. (2010). An analysis of field-aged diesel particulate filter performance: Particle emissions before, during, and after regeneration. *Journal of the Air & Waste Management Association*, 60(8):968–976.

Beatty, T. and Shimshack, J. (2011). School buses, diesel emissions, and respiratory health. *Journal of Health Economics*, 30(5):987–999.

Beatty, T. K. and Shimshack, J. P. (2014). Air pollution and children's respiratory health: A cohort analysis. *Journal of Environmental Economics and Management*, 67(1):39–57.

Biswas, S., Verma, V., Schauer, J. J., and Sioutas, C. (2009). Chemical speciation of pm emissions from heavy-duty diesel vehicles equipped with diesel particulate filter (dpf) and selective catalytic reduction (scr) retrofits. *Atmospheric Environment*, 43(11):1917 – 1925.

Bjornson, C. and Mitchell, I. (2000). Gender differences in asthma in childhood and adolescence. *Journal of Gender Specific Medicine*, 3(8):57 – 61.

Blasingame, K. (2012). Measurement agreement of fitnessgram aerobic capacity and body composition standards. *Iowa State University Graduate Theses and Dissertations*.

Borak, J. and Sirianni, G. (2007). Studies of self-pollution in diesel school buses: methodological issues. *Journal of occupational and environmental hygiene*, 4(9):660–668.

Calderón-Garcidueñas, L., Mora-Tiscareño, A., Styner, M., Gómez-garza, G., Zhu, H., Torres-Jardón, R., Carlos, E., Solorio-López, E., Medina-Cortina, H., Kavanaugh, M., and D'Angiulli, A. (2012). White matter hyperintensities, systemic inflammation, brain growth, and cognitive functions in children exposed to air pollution. *Journal of Alzheimer's Disease*, 31(1):183–191.

Castelli, D. M., Hillman, C. H., Buck, S. M., and Erwin, H. E. (2007). Physical fitness and academic achievement in third-and fifth-grade students. *Journal of Sport and Exercise Psychology*, 29(2):239–252.

Castillo-Garzón, M. J., Ruiz, J., Ortega, F. B., and Gutiérrez, A. (2006). Anti-aging therapy through fitness enhancement. 1:213–20.

Castro-Piñero, J., Perez-Bey, A., Segura-Jiménez, V., Aparicio, V. A., Gómez-Martínez, S., Izquierdo-Gomez, R., Marcos, A., Ruiz, J. R., Marcos, A., Castro-Piñero, J., et al. (2017). Cardiorespiratory fitness cutoff points for early detection of present and future cardiovascular risk in children: a 2-year follow-up study. In *Mayo Clinic Proceedings*, volume 92, pages 1753–1762. Elsevier.

Chen, X., Zhang, X., and Zhang, X. (2017). Smog in Our Brains: Gender Differences in the Impact of Exposure to Air Pollution on Cognitive Performance. *GLO Discussion Paper Series*, (32).

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, 126(4):1593–1660.

Clougherty, J. E. and Kubzansky, L. D. (2008). Traffic-related air pollution and stress: Effects on asthma. *Environmental Health Perspectives*, 116(9):A376–A377.

Commins, B. T., Waller, R. E., and Lawther, P. J. (1957). Air pollution in diesel bus garages. *British Journal of Industrial Medicine*, 14(4):232–239.

Currie, J., Hanushek, E. A., Kahn, E. M., Neidell, M., and Rivkin, S. G. (2009). Does Pollution Increase School Absences? *The Review of Economics and Statistics*, 91(4):682–694.

Currie, J. and Neidell, M. (2005). Air pollution and infant health: What can we learn from california's recent experience?*. *The Quarterly Journal of Economics*, 120(3):1003–1030.

Ebenstein, A., Lavy, V., and Roth, S. (2016). The long-run economic consequences of high-stakes examinations: Evidence from transitory variation in pollution. *American Economic Journal: Applied Economics*, 8(4):36–65.

Edwards, J. U., Mauch, L., and Winkelman, M. R. (2011). Relationship of nutrition and physical activity behaviors and fitness measures to academic performance for sixth graders in a midwest city school district. *Journal of School Health*, 81(2):65–73.

EPA (2003). Technical highlights: Questions and answers on using a diesel particulate matter filter in heavy-duty trucks and buses. Technical report, Environmental Protection Agency Office of Transportation and Air Quality.

Fahlman, M. M., Hall, H. L., and Lock, R. (2006). Ethnic and socioeconomic comparisons of fitness, activity levels, and barriers to exercise in high school females. *Journal of School Health*, 76(1):12–17.

Freire, C., Ramos, R., Puertas, R., Lopez-Espinosa, M.-J., Julvez, J., Aguilera, I., Cruz, F., Fernandez, M.-F., Sunyer, J., and Olea, N. (2010). Association of traffic-related air pollution with cognitive development in children. *Journal of Epidemiology & Community Health*, 64(3):223–228.

Gauderman, W. J., Avol, E., Lurmann, F., Kuenzli, N., Gilliland, F., Peters, J., and McConnell, R. (2005). Childhood asthma and exposure to traffic and nitrogen dioxide. *Epidemiology*, 16(6):737–743.

Gendron-Carrier, N., Gonzalez-Navarro, M., Polloni, S., and Turner, M. A. (2018). Subways and Urban Air Pollution. NBER Working Papers 24183, National Bureau of Economic Research, Inc.

Guxens, M. and Sunyer, J. (2012). A review of epidemiological studies on neuropsychological effects of air pollution. *The European Journal of Medical Sciences*, 141(3):1–7.

Ham, J. C., Zweig, J. S., and Avol, E. (2014). Pollution, test scores and the distribution of academic achievement: Evidence from california schools 2002-2008. *Manuscript, University of Maryland*.

Hammond, D., M. Lalor, M., and Jones, S. (2007). In-vehicle measurement of particle number concentrations on school buses equipped with diesel retrofits. *Water, Air, and Soil Pollution*, 179:217–225.

Harder, B. (2005). School buses spew pollution into young lungs. *Science News*, 167(21):334–334.

Hu, S., Herner, J. D., Shafer, M., Robertson, W., Schauer, J. J., Dwyer, H., Collins, J., Huai, T., and Ayala, A. (2009). Metals emitted from heavy-duty diesel vehicles equipped with advanced pm and nox emission controls. *Atmospheric Environment*, 43(18):2950 – 2959.

Institute of Medicine (2012). *Fitness Measures and Health Outcomes in Youth*. The National Academies Press, Washington, DC.

Jiang, Y., Yang, J., Cocker, D., Karavalakis, G., Johnson, K. C., and Durbin, T. D. (2018). Characterizing emission rates of regulated pollutants from model year 2012+ heavy-duty diesel vehicles equipped with dpf and scr systems. *Science of The Total Environment*, 619-620:765 – 771.

Kodama, S., Saito, K., Tanaka, S., Maki, M., Yachi, Y., Asumi, M., Sugawara, A., Totsuka, K., Shimano, H., Ohashi, Y., et al. (2009). Cardiorespiratory fitness as a quantitative predictor of all-cause mortality and cardiovascular events in healthy men and women: a meta-analysis. *JAMA*, 301(19):2024–2035.

Lakoski, S. G., Willis, B. L., Barlow, C. E., Leonard, D., Gao, A., Radford, N. B., Farrell, S. W., Douglas, P. S., Berry, J. D., DeFina, L. F., et al. (2015). Midlife cardiorespiratory fitness, incident cancer, and survival after cancer in men: the cooper center longitudinal study. *JAMA oncology*, 1(2):231–237.

Lavy, V., Ebenstein, A., and Roth, S. (2014). The impact of short term exposure to ambient air pollution on cognitive performance and human capital formation. Working Paper 20648, National Bureau of Economic Research.

Li, F., Lee, E. S., Liu, J., and Zhu, Y. (2015). Predicting self-pollution inside school buses using a cfd and multi-zone coupled model. *Atmospheric Environment*, 107:16–23.

Marshall, J. D. and Behrentz, E. (2005). Vehicle self-pollution intake fraction: children's exposure to school bus emissions. *Environmental Science & Technology*, 39(8):2559–2563. PMID: 15884349.

Monahan, P. (2006). School Bus Pollution Report Card 2006 : Grading the States. Report, Union of Concerned Scientists.

Murray, T., Eldridge, J., Silvius, P., Silvius, E., and Squires, W. G. (2012). Fitnessgram® friday: A middle school physical activity and fitness intervention. internationa. *Journal of Exercise Science*, 5(1):4–15.

Muzyka, V., Veimer, S., and Shmidt, N. (1998). Particle-bound benzene from diesel engine exhaust. *Scandinavian Journal of Work, Environment Health*, 24(6):481–485.

Ngo, N. S. (2015). Analyzing the relationship between bus pollution policies and morbidity using a quasi-experiment. *Medicine*, 94(37).

Ngo, N. S. (2017). Emission standards, public transit, and infant health. *Journal of policy analysis and management*, 36 4:773–89.

Ortega, F., Ruiz, J., Castillo, M., and Sjöström, M. (2008). Physical fitness in childhood and adolescence: a powerful marker of health. *International journal of obesity*, 32(1):1.

Reichardt, R. (2000). *The Cost of Class Size Reduction: Advice for Policy Makers*. Ph.D. dissertation, RAND Corporation.

Rice, J. K. (2010). The impact of teacher experience examining the evidence and policy implications. *CALDER*, Brief 11.

Ross, R., Blair, S. N., Arena, R., Church, T. S., Després, J.-P., Franklin, B. A., Haskell, W. L., Kaminsky, L. A., Levine, B. D., Lavie, C. J., Myers, J., Niebauer, J., Sallis, R., Sawada, S. S., Sui, X., and Wisløff, U. (2016). Importance of assessing cardiorespiratory fitness in clinical practice: A case for fitness as a clinical vital sign. *AHA*.

Sappok, A., Santiago, M., Vianna, T., and Wong, V. (2009). Characteristics and effects of ash accumulation on diesel particulate filter performance: Rapidly aged and field aged results. *SAE Technical Paper*.

Soll-Johanning, H., Bach, E., Olsen, J. H., and Tüchsen, F. (1998). Cancer incidence in urban bus drivers and tramway employees: A retrospective cohort study. *Occupational and Environmental Medicine*, 55(9):594–598.

Stevens, T., Cheng, W., Jaspers, I., and Madden, M. (2010). Effect of short-term exposure to diesel exhaust particles and carboxylic acids on mitochondrial membrane disruption in airway epithelial cells. 181:A1031–A1031.

Sunyer, J., Esnaola, M., Alvarez-Pedrerol, M., Forns, J., Rivas, I., López-Vicente, M., Suades-González, E., Foraster, M., Garcia-Esteban, R., Basagaña, X., Viana, M., Cirach, M., Moreno, T., Alastuey, A., Sebastian-Galles, N., Nieuwenhuijsen, M., and Querol, X. (2015). Association between traffic-related air pollution in schools and cognitive development in primary school children: A prospective cohort study. *PLOS Medicine*, 12(3):1–24.

Tate, J., Mason, R., and Schmitt, L. (2017). 2050 - the air quality emissions and health benefits of cleaner buses: A city of york (uk) case study using micro-scale models and a health impact toolkit. *Journal of Transport Health*, 5:S41.

Waller, R. E., Hampton, L., and Lawther, P. J. (1985). A further study of air pollution in diesel bus garages. *British Journal of Industrial Medicine*, 42(12):824–830.

Welk, G. J., Jackson, A. W., Jr., J. R. M., Haskell, W. H., Meredith, M. D., and Cooper, K. H. (2010). The association of health-related fitness with indicators of academic performance in texas schools. *Research Quarterly for Exercise and Sport*, 81(sup3):S16–S23. PMID: 21049834.

Xu, W., Mai, G., Zhu, Q., Yu, Z., and Liu, Y. (2016). Pollution exposure at bus commuter stations in guangzhou, china. *International Journal of Environmental Technology and Management*, 19(2):103–119.

Zhang, Q. and Zhu, Y. (2011). Performance of school bus retrofit systems: ultrafine particles and other vehicular pollutants. *Environmental science & technology*, 45(15):6475–6482.

Zuurbier, M., Hoek, G., Oldenwening, M., Lenters, V., Meliefste, K., van den Hazel, P., and Brunekreef, B. (2010). Commuters' exposure to particulate matter air pollution is affected by mode of transport, fuel type, and route. *Environmental Health Perspectives*, 118(6):783–789.

Tables

Table 1: District-Level Student Characteristics

	(1) Non-Retrofitting Districts		(2) Retrofitting Districts		(3) Difference T-Test of Means	
Health Outcomes (2012-2017)						
Aerobic Capacity ($V0_2$ Max)	41.160	(1.688)	41.201	(1.422)	-0.0412	(-0.12)
Body-Mass Index	21.069	(0.880)	20.633	(0.340)	0.436*	(2.54)
AC Attempts / Enrollment	0.407	(0.114)	0.425	(0.079)	-0.0174	(-0.76)
BMI Attempts / Enrollment	0.654	(0.153)	0.689	(0.108)	-0.0346	(-1.12)
Schooling Outcomes (2007-2017)						
Math Z-Scores	-0.107	(0.263)	-0.060	(0.216)	-0.0473	(-0.88)
ELA Z-Scores	-0.107	(0.229)	-0.061	(0.194)	-0.0459	(-0.98)
Attendance rate	95.573	(0.630)	95.584	(0.488)	-0.0112	(-0.09)
Demographics (2007-2017)						
African American	0.367	(0.272)	0.363	(0.266)	0.004	(0.07)
Hispanic	0.082	(0.105)	0.109	(0.077)	-0.028	(-1.32)
White	0.554	(0.252)	0.504	(0.276)	0.051	(0.95)
Other	0.030	(0.025)	0.055	(0.029)	-0.025***	(-4.71)
Male	0.513	(0.010)	0.513	(0.005)	0.000	(0.11)
Female	0.487	(0.010)	0.487	(0.005)	-0.000	(-0.11)
Students (thousands)	5.655	(9.765)	28.081	(37.502)	-22.426***	(-6.34)
Free and Reduced Lunch	0.668	(0.171)	0.616	(0.146)	0.052	(1.49)
Students with Disabilities	0.123	(0.024)	0.121	(0.018)	0.002	(0.38)
English Language Learner	0.025	(0.037)	0.045	(0.043)	-0.021*	(-2.58)
Retrofits (2007-2017)						
Buses Retrofitted per Retrofit			66.39	(145.3)		
Proportion of Fleet Retrofitted			0.189	(0.141)		
Average Retrofit Cost per Bus ($)			8111.0	(5013.8)		
Bus Fleet Characteristics (2007-2017)						
Average Time in Bus (minutes)	44.883	(11.629)	49.631	(7.940)	-4.748*	(-2.04)
District Bus Ridership	0.621	(0.174)	0.610	(0.087)	0.0113	(0.33)
Total Buses	75.594	(104.432)	313.286	(411.857)	-237.7***	(-6.17)
Total Bus Riders (thousands)	3.475	(7.412)	17.563	(25.814)	-14.088***	(-5.62)
Average Bus Age	14.126	(1.574)	14.268	(1.537)	-0.142	(-0.43)
Observations	153		27		180	

Mean coefficients reported; standard deviations in parentheses. Observations are at the district level. Other demographic category includes Asian, American Indian, Pacific Islander, and Multiracial. Students represents the average student enrollment in thousands. Standardized math and ELA test scores are negative because the majority of Georgia school districts are rural, small, and under-achieving relative to larger urban districts. Aerobic capacity attempts / enrollment represents the number of attempts divided by K-12 enrollment, where certain grades in high school are never tested on the FitnessGram examination.

Table 2: FitnessGram Health 2012-2017

	(1) AC	(2) AC	(3) AC	(4) BMI	(5) BMI	(6) BMI
Percent Retrofitted	1.815**			-0.241		
	(0.81)			(0.33)		
Percent Retrofitted Ridership		2.439*			-0.479	
		(1.36)			(0.53)	
Percent Retrofitted Ridership * Trip Duration			0.041			-0.010
			(0.03)			(0.01)
Dep. Var. mean	41.66	41.66	41.66	21.03	21.03	21.03
R2	0.197	0.199	0.198	0.050	0.051	0.051
N	856	846	846	863	853	853

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors at the district level in parentheses. Year fixed effects included. Demographic variables include the proportion of students that are Asian, African-American, Hispanic, and male, where White and female are the omitted categories, as well as the percentage of students with free or reduced price lunch, disabilities, and English-language learner status. Bus characteristics include average bus age, the proportion of buses built before 2007, and the proportion of liquid natural gas-, butane-, and gasoline-powered buses in the district. Mean-centered ridership and trip duration variables also included as controls. The independent variable percent retrofitted is the proportion of a district's bus fleet that is retrofitted in a given year, and zero otherwise. Percent retrofitted * ridership is the percent of the bus fleet retrofitted times the time-constant proportion of students in a district riding the bus, while percent retrofitted * ridership * trip duration is the proportion of the bus fleet retrofitted times time-constant ridership and the time-constant average duration of a daily bus commute for students in a given district. All-district mean is 41.66 for aerobic capacity and 21.03 for BMI.

Table 3: FitnessGram Health by Gender and School Type 2012-2017

	Elementary		Middle		High School	
	Male	Female	Male	Female	Male	Female
	(1)	(2)	(3)	(4)	(5)	(6)
Percent Retrofitted	3.963**	4.152*	-1.651	0.304	1.899**	1.802
	(1.99)	(2.19)	(1.26)	(2.02)	(0.78)	(1.41)
Dep. Var. mean	42.45	39.82	43.22	38.85	43.45	37.75
R^2	0.143	0.273	0.093	0.305	0.031	0.086
N	777	777	770	770	710	710

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors at the district level in parentheses. Outcomes are district average aerobic capacity among elementary schools only. Year fixed effects included. Demographic variables include the proportion of students that are Asian, African-American, Hispanic, and male, where White and female are the omitted categories, as well as the percentage of students with free or reduced price lunch, disabilities, and English-language learner status. Bus characteristics include average bus age, the proportion of buses built before 2007, and the proportion of liquid natural gas-, butane-, and gasoline-powered buses in the district. Mean-centered ridership and trip duration variables also included as controls. The independent variable percent retrofitted is the proportion of a district's bus fleet that is retrofitted in a given year, and zero else.

Table 4: Academic Achievement 2007-2017

	(1) ELA	(2) ELA	(3) ELA	(4) Math	(5) Math	(6) Math	(7) Attend	(8) Attend	(9) Attend
Percent Retrofitted	0.088***			0.049			0.136		
	(0.02)			(0.03)			(0.24)		
Percent Retrofitted		0.143***			0.083			0.259	
Ridership		(0.04)			(0.05)			(0.39)	
Percent Retrofitted			0.003***			0.001			0.006
Ridership * Trip Duration			(0.00)			(0.00)			(0.01)
Dep. Var. mean	-0.100	-0.100	-0.100	-0.099	-0.099	-0.099	95.57	95.57	95.57
R2	0.062	0.066	0.066	0.020	0.021	0.021	0.100	0.100	0.100
N	1,260	1,246	1,246	1,260	1,246	1,246	1,260	1,246	1,246

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors at the district level in parentheses. Year fixed effects included. Demographic control variables include the proportion of students that are Asian, African-American, Hispanic, and male, where White and female are the omitted categories, as well as the percentage of students with free or reduced price lunch, disabilities, and English-language learner status. Bus control variables include average bus age, the proportion of buses built before 2007, and the proportion of liquid natural gas-, butane-, and gasoline-powered buses in the district. Mean-centered ridership and trip duration variables also included as controls. The independent variable percent retrofitted is the proportion of a district's bus fleet that is retrofitted in a given year, and zero else. Percent retrofitted * ridership is the percent of the bus fleet retrofitted times the time-constant proportion of students in a district riding the bus, while percent retrofitted * ridership * trip duration is the proportion of the bus fleet retrofitted times time-constant ridership and the time-constant average duration of a daily bus commute for students in a given district.

Table 5: Academic Achievement by School Type 2007-2017

	Elementary		Middle	
	ELA	Math	ELA	Math
	(1)	(2)	(3)	(4)
Percent Retrofitted	0.119***	0.061	0.059**	0.047
	(0.03)	(0.07)	(0.03)	(0.03)
Dep. Var. mean	-0.091	-0.089	-0.107	-0.107
R^2	0.043	0.02	0.042	0.037
N	1,800	1,800	1,800	1,800

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors at the district level in parentheses. Year fixed effects included. Demographic variables include the proportion of students that are Asian, African-American, Hispanic, and male, where White and female are the omitted categories, as well as the percentage of students with free or reduced price lunch, disabilities, and English-language learner status. Bus characteristics include average bus age, the proportion of buses built before 2007, and the proportion of liquid natural gas-, butane-, and gasoline-powered buses in the district. Mean-centered ridership and trip duration variables also included as controls. The independent variable percent retrofitted is the proportion of a district's bus fleet that is retrofitted in a given year, and zero else. Elementary includes end-of-grade test scores for grades 3-5, while middle includes the same for grades 6-8.

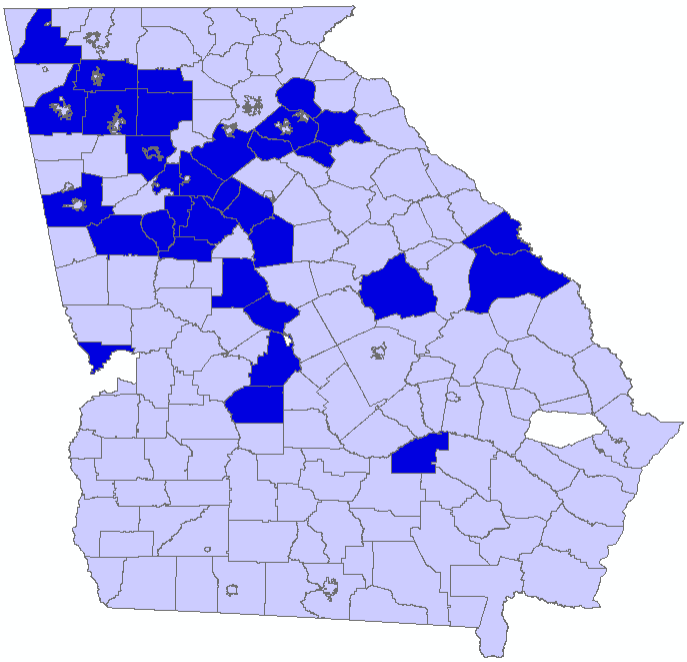
Table 6: All Outcomes by Retrofit Type

	(1) ELA	(2) Math	(3) Attend	(4) AC	(5) BMI
Diesel Particulate Filter	0.134**	0.063	0.459	1.411	-0.612
	(0.05)	(0.07)	(0.52)	(1.89)	(0.54)
Closed-Crankcase Filter	-0.022	-0.012	-0.635	-	-
	(0.04)	(0.05)	(0.45)	(.)	(.)
Diesel Oxidation Catalyst	0.051**	0.047	0.144	1.367	-0.139
	(0.02)	(0.03)	(0.19)	(0.85)	(0.46)
Flow-through Filter	-0.026	-0.177***	-0.149	-	-
	(0.06)	(0.05)	(1.43)	(.)	(.)
Dep. Var. mean	-0.100	-0.099	95.57	41.66	21.03
R^2	0.058	0.023	0.096	0.186	0.049
N	1,800	1,800	1,800	856	863

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors at the district level in parentheses. Year fixed effects included. Demographic control variables include the proportion of students that are Asian, African-American, Hispanic, and male, where White and female are the omitted categories, as well as the percentage of students with free or reduced price lunch, disabilities, and English-language learner status. Mean-centered ridership and trip duration variables also included as controls. The number of buses replaced early also included as a control. Bus characteristics not included due to high correlation with covariates. The independent variables each represent the proportion of a bus fleet that is retrofitted with the given engine modification. The sample includes 32 DPF retrofits, nine CCF retrofits, eight DOC retrofits, and three flow-through filter retrofits. Accordingly, results for flow-through filter retrofits may be unreliable.

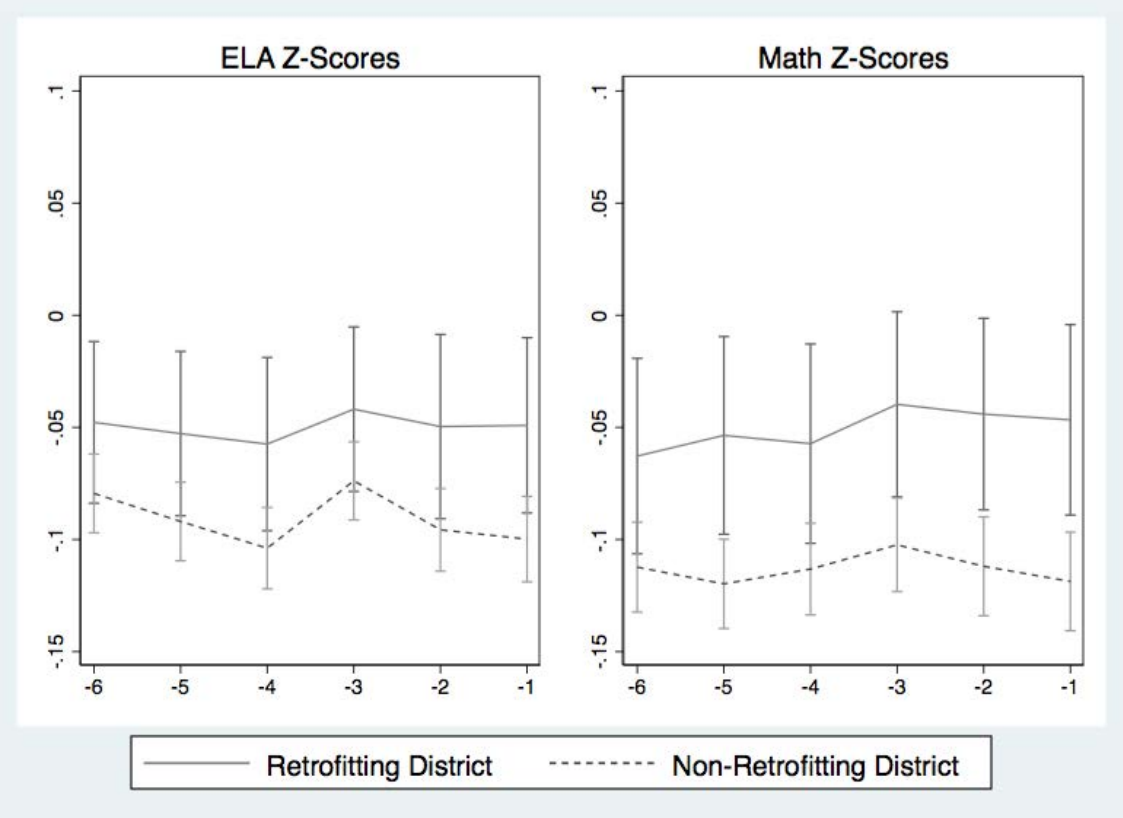
Figures

Figure 1: Retrofitting School Districts



Notes: Darker blue school districts have at least one retrofit cycle during the relevant sample window (2007-2017). Blank districts are missing data.

Figure 2: Differential Pre-Trends by Retrofitting Districts 2007-2013



Notes: Figure plots the trend in ELA and Math test scores across retrofitting and non-retrofitting districts before 2013, the mode year of retrofit implementation, such that -1 represents school year 2011-12. Because the timing of retrofits varied across districts, we are unable to conduct a simple event study with non-retrofitting districts as a comparison. We therefore normalize treatment to 2013 and plot trends across districts that ever retrofit and those that do not. Some of the pre-2013 years feature retrofits, and therefore may be expected to have differential trends over this period. Nevertheless, the trend lines are roughly parallel over this sample window.

Appendix

Table A1: Academic Achievement by Grade 2007-2017

	Grade 3/6		Grade 4/7		Grade 5/8	
	ELA	Math	ELA	Math	ELA	Math
	(1)	(2)	(3)	(4)	(5)	(6)
I. Elementary Schools						
Percent Retrofitted	0.087	0.034	0.208**	0.203*	0.169***	0.060
	(0.07)	(0.11)	(0.10)	(0.11)	(0.05)	(0.08)
R^2	0.033	0.022	0.072	0.064	0.056	0.069
II. Middle Schools						
Percent Retrofitted	0.048	-0.003	0.049	0.031	0.073*	0.108
	(0.05)	(0.06)	(0.05)	(0.04)	(0.04)	(0.08)
R^2	0.048	0.015	0.037	0.024	0.016	0.038
N	1,800	1,800	1,800	1,800	1,800	1,800

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors at the district level in parentheses. Outcomes are grade-level ELA and math scores. Year fixed effects included. Demographic variables include the proportion of students that are Asian, African-American, Hispanic, and male, where White and female are the omitted categories, as well as the percentage of students with free or reduced price lunch, disabilities, and English-language learner status. Bus characteristics include average bus age, the proportion of buses built before 2007, and the proportion of liquid natural gas-, butane-, and gasoline-powered buses in the district. Mean-centered ridership and trip duration variables also included as controls. The independent variables percent retrofitted is the proportion of a district's bus fleet that is retrofitted in a given year, and zero else.

Table A2: Sensitivity of Aerobic Capacity Results to Different Cutoffs 2012-2017

	(1) 15	(2) 20	(3) 25	(4) 30	(5) 35	(6) 30 & 26	(7) Jumps	(8) None	(9) 2012
Percent Retrofitted	2.298	1.313	1.483	1.763**	1.675**	1.815**	3.528***	1.324	7.089***
	(2.07)	(1.34)	(0.97)	(0.73)	(0.70)	(0.81)	(0.85)	(2.29)	(1.09)
R^2	0.248	0.238	0.223	0.218	0.147	0.197	0.098	0.246	0.300
N	860	860	860	857	849	856	675	860	681

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors at the district level in parentheses. Year fixed effects included. Demographic and bus characteristics included as controls. Mean-centered ridership and trip duration variables also included as controls. Column headers represent different $V0_2$ max cutoff values. Average aerobic capacity in 2011-12 and 2012-13 is left-skewed, with many implausibly low values for $V0_2Max$. In later years, no school-average $V0_2Max$ is below 30 for male assessments and 26 for female assessments. We therefore demonstrate aerobic capacity results under a range of cutoffs, where each cutoff represents dropping school-level aerobic capacity results below the given value. In column (7), labeled Jumps, we replace as missing any school with average values that increase or decrease by more than 6 $V0_2$ max units from 2011-12 to 2012-13. These jumps are very large in relation to those observed after 2012-13, and so dropping these observations is often equivalent to dropping all values below a given low-valued cutoff. In the column (2012), we drop the entire year of 2011-12, which restricts the number of retrofitting districts such that the coefficient is estimated from only three retrofitting districts. We prefer model (6), the cutoff at 30 for males and 26 for females, because it creates a 2012 distribution that best conforms to the other years of the sample while simultaneously not dropping too many low yet accurate results. In almost all cases, the cutoffs drop less than a tenth of school observations in any given district. Controlling for the proportion of schools dropped does not affect the results because the proportion dropped is not correlated with treatment.

Table A3: Correlation of Proportion of a Bus Fleet Retrofitted with District Characteristics 2007-2017

	(1) ΔAC Part.	(2) ΔBMI Part.	(3) ΔRidership
I. Endogenous Responses			
Percent Retrofitted	-0.449	-0.576*	-0.023
	(0.32)	(0.30)	(0.04)
R^2	0.153	0.030	0.012
N	870	870	1,780
	Δ Bus Age	Δ Total Buses	Δ Trip Duration
II. Bus Characteristics			
Percent Retrofitted	0.470	56.662	4.227
	(0.63)	(36.02)	(5.80)
R2	0.129	0.057	0.011
N	1,800	1,800	1,780
	Δ Afr. American	Δ Hispanic	Δ Male
III. Student Demographics			
Percent Retrofitted	0.015	0.466	-0.109
	(0.36)	(0.38)	(0.30)
R2	0.021	0.028	0.007
N	1,800	1,800	1,800
	Δ ELL	Δ SWD	Δ FRPL
IV. Student Characteristics			
Percent Retrofitted	-0.009	0.572**	0.924
	(0.08)	(0.29)	(1.58)
R2	0.012	0.219	0.035
N	1,800	1,800	1,800

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors at the district level in parentheses. Year fixed effects included. In the first panel, models (1) and (2) demonstrate the extent to which the proportion of a bus fleet retrofitted is correlated with changes in the participation rate, i.e. the number of attempts divided by district enrollment. Model (3) shows whether the proportion of a bus fleet retrofitted is correlated with year-on-year changes the ridership rate. The relevant sample for models (1) and (2) is 2012 - 2017, while model (3) covers the entire sample window, 2007-2017. In the second panel, we show that the proportion of a bus fleet retrofitted is not significantly correlated with changes in the average bus age within a district, the total number of buses, or the average trip duration. The third panel demonstrates that the proportion of a bus fleet retrofitted is not significantly related to changes in the percent of a district that is African American, Hispanic, or Male. The fourth panel shows the relationship between the proportion of a bus fleet retrofitted and changes in the percent of a district's students that are English language learner, students with disabilities, or receiving free- and reduced-price lunch.

Table A4: Academic Achievement 2007-2017, Dropping Milestones Years 2015-2017

	(1) ELA	(2) ELA	(3) ELA	(4) Math	(5) Math	(6) Math
Percent Retrofitted	0.089***			0.049		
	(0.03)			(0.03)		
Percent Retrofitted Ridership		0.143***			0.083*	
		(0.04)			(0.05)	
Percent Retrofitted Ridership * Trip Duration			0.002***			0.001
			(0.00)			(0.00)
R^2	0.064	0.064	0.064	0.020	0.020	0.020
N	1,440	1,440	1,440	1,440	1,440	1,440

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors at the district level in parentheses. Outcomes are district average ELA test scores, Math test scores, attendance, aerobic capacity, and BMI. District and year fixed effects included. Demographic variables include the proportion of students that are Asian, African-American, Hispanic, and male, where White and female are the omitted categories, as well as the percentage of students with free or reduced price lunch, disabilities, and English-language learner status. Bus characteristics include average bus age, the proportion of buses built before 2007, and the proportion of liquid natural gas-, butane-, and gasoline-powered buses in the district. Controls for ridership share and trip duration are also included. The table shows how our first-differences estimates change when dropping all years after 2014-15 when the new Milestones standardized examination is offered instead of the CRCT exam. Milestones computerized examinations suffered from widespread glitches that may have affected our estimates.

Table A5: Drop Interpolated Bus Years

	(1) ELA	(2) Math	(3) ATT	(4) AC	(5) BMI
Percent Retrofitted	0.083***	0.057	0.242	1.766**	-0.242
	(0.03)	(0.04)	(0.29)	(0.83)	(0.35)
R2	0.079	0.029	0.174	0.188	0.061
N	1,260	1,260	1,260	692	698

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors at the district level in parentheses. Outcomes are district average ELA test scores, Math test scores, attendance, aerobic capacity, and BMI. District and year fixed effects included. Demographic variables include the proportion of students that are Asian, African-American, Hispanic, and male, where White and female are the omitted categories, as well as the percentage of students with free or reduced price lunch, disabilities, and English-language learner status. Bus characteristics include average bus age, the proportion of buses built before 2007, and the proportion of liquid natural gas-, butane-, and gasoline-powered buses in the district. Controls for ridership share and trip duration are also included. The independent variable percent retrofitted is the proportion of a district's bus fleet that is retrofitted in a given year, and zero else. The table shows how our first-differences estimates change when dropping all years for which information on district bus fleets is lacking. For these years, we inserted the value of the nearest year for which data is available, which is 2010 for all years prior and 2016 for 2017.

Table A6: Timing of Retrofit Implementation

	(1) ELA	(2) Math	(3) Attend	(4) AC	(5) BMI
I. Regular Timing					
Percent Retrofitted	0.089***	0.049	0.154	1.815**	-0.241
	(0.03)	(0.03)	(0.25)	(0.81)	(0.33)
R^2	0.058	0.023	0.097	0.197	0.050
N	1,800	1,800	1,800	856	863
II. Treatment 1-year in Advance					
Percent Retrofitted	-0.029	-0.040	-0.110	2.258	0.197
	(0.03)	(0.04)	(0.23)	(2.75)	(0.62)
R^2	0.056	0.023	0.097	0.196	0.050
N	1,800	1,800	1,800	856	863
III. January Implementation					
Percent Retrofitted	0.100***	0.079**	0.273	1.664*	-0.089
	(0.02)	(0.03)	(0.26)	(0.87)	(0.28)
R^2	0.059	0.024	0.097	0.197	0.050
N	1,800	1,800	1,800	856	863

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors in parentheses. Year fixed effects included. Demographic control variables include the proportion of students that are Asian, African-American, Hispanic, and male, where White and female are the omitted categories. The percentage of students with free or reduced price lunch, disabilities, and English-language learner status. Bus characteristics include average bus age, the proportion of buses built before 2007, and the proportion of liquid natural gas-, butane-, and gasoline-powered buses in the district. Mean-centered ridership and trip duration variables also included as controls. The independent variable percent retrofitted varies according to the timing of the retrofit completion date in a given school district. In the base case presented in Panel I, regular timing, all retrofits between May and the following April are assigned to the fiscal year of the latter April. In Panel II, we show the results when assigning a placebo treatment year as one year before the actual retrofit completion year. In Panel III, retrofits completed before January are assigned to the same fiscal year, but those occurring after January are assigned to the following fiscal year.

Table A7: All Outcomes Fixed Effects Estimates

	(1) ELA	(2) Math	(3) Attend	(4) AC	(5) BMI
I: First Differences					
Percent Retrofitted	0.089***	0.049	0.154	1.815**	-0.241
	(0.03)	(0.03)	(0.25)	(0.81)	(0.33)
R^2	0.058	0.023	0.097	0.197	0.050
N	1,800	1,800	1,800	856	863
II: Fixed Effects					
Percent Retrofitted	0.092	-0.006	0.130	1.108	-0.166
	(0.06)	(0.08)	(0.22)	(1.61)	(0.34)
R^2	0.900	0.900	0.391	0.705	0.613
N	1,958	1,958	1,958	1,034	1,040
D-W F-Stat	197.77	194.42	17.01	31.68	11.39
Prob > F	0.0000	0.0000	0.0001	0.0000	0.0009
III: FE Adding Leads					
Percent Retrofitted	0.092**	0.038	0.043	0.363	-0.410
	(0.04)	(0.06)	(0.31)	(1.51)	(0.38)
Percent Retrofit Lead	-0.011	-0.050	0.205	5.398	-0.254
	(0.05)	(0.07)	(0.29)	(3.55)	(0.72)
N	1800	1800	1800	877	882

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors at the district level in parentheses. Year and district fixed effects included. Outcomes are district average ELA test scores, Math test scores, attendance, aerobic capacity, and BMI. District and year fixed effects included. Demographic variables include the proportion of students that are Asian, African-American, Hispanic, and male, where White and female are the omitted categories, as well as the percentage of students with free or reduced price lunch, disabilities, and English-language learner status. Bus characteristics include average bus age, the proportion of buses built before 2007, and the proportion of liquid natural gas-, butane-, and gasoline-powered buses in the district. Mean-centered ridership and trip duration variables also included as controls. The independent variable percent retrofitted is the proportion of a district's bus fleet that has ever been retrofitted, and zero else. The table estimates all outcomes using the fixed-effects model. It also compares our estimates on test scores when we drop the problematic Milestones years 2015-16 and 2016-17 when widespread technical difficulties with the new computer-based testing system caused the state to throw out the results for accountability purposes. All-district mean is 41.66 for aerobic capacity and 21.03 for BMI.

Figure A1: Data Issues in Aerobic Capacity

Notes: Each pane scatters the school-level average VO_2 max against the percent of students attaining healthy fitness zone (HFZ) status. The left pane presents the scatterplot for school years 2011-12 and 2012-13, while the right pane displays a scatterplot for the remaining years in the sample. A school's average VO_2 max should be highly correlated with the percent of students attaining HFZ status because each child's VO_2 max is used to determine whether they meet HFZ standards. In the right panel, we observe such a tight relationship between these related measures. In the left panel, however, the relationship is less clear. After the 2012-13 school year, there are no female school-level VO_2 max observations below 26 or male school-level VO_2 max observations below 30. These values are nevertheless very common in the first two years of the sample, and many of these low values correspond to relatively high HFZ attainment. Such values suggest a data-reporting issue in the roll-out years of the sample.